From Reasons to Causes:

A Theory of Causation

Holger Andreas[1] and Mario Günther[2]

[1]University of British Columbia
[2]LMU Munich and Carnegie Mellon University

Cambridge University Press

To Laurenz – H. A.

To Atoosa, Maria, and Konrad – M. G.

Preface

From reasons to causes is the leitmotif of this book. The two authors came together for the first time to explore belief revision theory and the Ramsey Test for an analysis of the conjunction 'because' in natural language. This was the starting point of the *epochetic* approach to causation to be developed here. In a series of papers, we have explored a Ramsey Test analysis of reasons for an analysis of causation. The theory of causation in this book builds on these papers, but goes well beyond them. Specifically, we take up the challenge to devise a reductive analysis of causation, which does not take any causal or modal notions for granted.

This book has many sources of inspiration. Two of them are particularly noteworthy. Hans Rott was the first to study variants of the Ramsey Test for an account of reasons in everyday and scientific contexts. Wolfgang Spohn came up with the idea to build a theory of causation on top of an analysis of reasons. We had the pleasure to meet Hans and Wolfgang in Munich for academic talks, which drew our attention to their work on said topics.

The Munich Center for Mathematical Philosophy at LMU Munich was the institutional starting point of this book. We are grateful to our teachers and mentors for the opportunity to spend many productive years there. In particular, we would like to thank Ulises Moulines, Hannes Leitgeb, Stephan Hartmann, and Christian List. We were fortunate to continue the project underlying this book in other parts of the world. Holger Andreas went to Canada to take up a professorship at the University of British Columbia. He is most grateful to the department heads Andrew Irivine, Helen Yanacopulos, and Noriko Ozawa for supporting his career and research there. His work was furthermore supported by grants from the German Research Council and Canada's Social Sciences and Humanities Research Council.

Upon completion of his PhD, Mario Günther moved to the Australian National University. He is most grateful to Alan Hájek, Katie Steele, Phil Dowe, and Philip Pettit for feedback on the book project, as well as general guidance and support. He then returned to the Munich Center for Mathematical Philosophy as an assistant professor. His work was supported by the Excellence Initiative of the German federal and state governments, a junior residency in the Center for Advanced Studies of LMU Munich, and a visiting assistant professorship awarded by Carnegie Mellon University.

The book benefitted greatly from the commentaries on our lead article 'Factual Difference-Making' forthcoming in the *Australasian Philosophical Review*. They made us see, for example, that a simple distinction between deviant and non-deviant events is insufficient to capture our causal judgments in some scenarios. We would like to extend special thanks to Cei Maslen for outstanding editorial support.

Part I of the book and the forthcoming lead article propose epochetic analyses of causation. One crucial difference is that the book analysis relies on the inferential notion of an *active path*, which represents how an effect depends on its cause in a causal process. This notion allows us to distinguish preempted from genuine causes even in the absence of events between them in the causal process. Another difference is that the book analysis does not exploit *indeterminate interventions*. But this difference is less crucial because the book analysis could be enriched by such interventions in the spirit of the epochetic analysis.

We are very grateful to Hilary Gaskin at Cambridge University Press for supporting this project. Special thanks are also due to two anonymous referees whose reports greatly helped us improve our theory of causation. Finally, we would like to thank each other.

Contents

Preface **ii**

1 Introduction **1**

1 The Epochetic Approach 2

2 The Concept of Causation 4

3 Reasons 8

4 Causes in Causal Models 9

5 The Reductive Analysis 12

6 Events 15

7 Advantages over the Counterfactual Approach 21

I Causes in Causal Models **25**

2 Active Paths **26**

1 Causal Models 27

2 Causal Graphs 38

3 Inferential Pathways 40

4 The Inferential Analysis 41

5 Inferential Networks 46

3 Classics 53
1 Neuron Diagrams 53
2 Conjunctive Causes 55
3 Overdetermination 58
4 Early Preemption 59
5 Direct Deductions and Genuine Causation 63
6 Late Preemption 65
7 Trumping Preemption 68
8 Prevention 71
9 Double Prevention 73

4 Non-Transitivity 76
1 Short Circuit 77
2 Extended Double Prevention 79
3 Simple Switch 83
4 Basic Switch 86
5 Realistic Switch 88
6 Sartorio's Principle 90

5 Entanglement 94
1 Subcause 94
2 What is Entanglement? 96
3 Collaboration 100
4 Failed Collaboration 102
5 Disjunctive Cause Follows Conjunctive Cause 103
6 Conjunctive Cause Follows Disjunctive Cause 104
7 Disjunctive Cause Opposes Conjunctive Cause 105
8 Conjunctive Cause Opposes Disjunctive Cause 106

6 Deviancy 108
1 The Problem of Isomorphic Causal Models 108
2 Deviancy and Normality 111
3 Deviancy of Causes 115
4 Bogus Prevention 117
5 Extended Bogus Prevention 118
6 Bogus Double Prevention 120
7 Extended Short Circuit 123
8 Modified Extended Double Prevention 124
9 Isomorphic Modified Extended Double Prevention 126
10 Omissions 128
11 Realistic Switches 131
12 Looking Back: Preservation of Causal Verdicts 134
13 Revisiting Overdetermination and Preemption 136

II The Reductive Analysis 141

7 A Humean Analysis 143
1 The Humean Convention 144
2 Poincarean Conventions 148
3 Belief Revision Theory 152
4 Epochetic Conditionals 157
5 Causation 159

8 Spurious Causation 164
1 Common Causes 164
2 Non-Redundant Laws 171
3 Abductive Inferences 172
4 Abductively Redundant Laws 179

5 Common Causes in Conjunctive Scenarios 180
6 Independence of Causes 184
7 Common Causes in Disjunctive Scenarios 185
8 Conjunctive and Disjunctive Scenarios Combined 189
9 Lightning 191
10 Trivial Spurious Causes 196
11 An Ordering of Unification 199
12 Unification and Causation 203
13 Other Solutions 207

9 Simultaneous Causation 214

1 Some Examples 214
2 Causal Explanatory Asymmetries 216
3 Causal Scenarios 219
4 The Length of the Pendulum 230

10 Backward Causation 234

1 From an Interventionist Perspective 235
2 The Interventionist Account with Causal Graphs 236
3 The Direction of Interventions 239
4 The Bilking Argument 244
5 Ways out of the Bilking Argument 246
6 The Original Fork Theory 251
7 The Counterfactual Approach 260
8 Problems of the Counterfactual Approach 265
9 The Independence Theory 274
10 The Disjunctive Fork Theory 279
11 Conclusion 284

11 Conclusion and Synthesis **288**

1 Causes in Causal Models 289

2 The Reductive Analysis 291

3 Synthesis 292

4 The Final Analysis 296

5 Causation in the Objects 302

Appendices **306**

A The Logic of Causal Models **306**

1 Motivation 307

2 Propositional Causal Models 308

3 Natural Deduction 310

4 Semantics 314

5 Soundness and Completeness 316

6 Non-Binary Variables 318

B Belief Revision Theory **324**

1 Revision of a Ranked Belief Base 324

2 Epochetic Conditionals 327

C Proofs **330**

1 Chapter 2 330

2 Chapter 3 338

3 Chapter 5 346

Bibliography **349**

Index of Names **362**

Index of Subjects **363**

List of Figures

1 Early preemption 30

2 Causal graph of early preemption 39

3 Conjunctive causes 56

4 Agnostic model for conjunctive causes 57

5 Active path from C to E 57

6 Overdetermination 58

7 Agnostic model for overdetermination 59

8 Active path from C to E 59

9 Agnostic model for early preemption 60

10 Active path from C to E 61

11 Late preemption 66

12 Agnostic model for late preemption 67

13 Active path from C to E 67

14 Trumping 69

15 Agnostic model for trumping 70

16 Active path from C to E 70

17 Prevention 72

18 Agnostic model for prevention 72

19 Active path from C to E 73

20 Double prevention 74
21 Agnostic model for double prevention 75
22 Active path from C to E 75

23 Short circuit 77
24 Causal model agnostic on D and E 79
25 Active path from D to $\neg E$ 79
26 Extended double prevention 80
27 Causal model agnostic on F and E 81
28 Causal model agnostic on C and E 82
29 Active path from C to E 82
30 Simple switch 84
31 Causal model agnostic on F and R 85
32 Active path from F to R 85
33 Causal model agnostic on R and E 86
34 Active path from R to E 86
35 Neuron diagram for basic switch 87
36 Dependency diagram for basic switch 87
37 Realistic switch 89
38 Extended early preemption 91

39 Subcause 95
40 Agnostic model for subcause 95
41 Active path from C to E 96
42 Collaboration 101
43 A variant of collaboration 103
44 Another variant of collaboration 104

45 Bogus prevention 109

45 Bogus prevention . . . 117

46 Causal model agnostic on F and $\neg E$ . . . 117

47 Active path from F to $\neg E$ . . . 118

48 Extended bogus prevention . . . 119

49 Bogus double prevention . . . 121

50 Agnostic model for bogus double prevention . . . 122

51 Active path from C to E . . . 122

52 Extended short circuit . . . 124

53 Modified extended double prevention . . . 125

54 Isomorphic modified extended double prevention . . . 126

55 Causal model agnostic on C and E . . . 127

56 Active path from C to E . . . 127

57 Omission . . . 128

58 Agnostic model for omission . . . 129

59 Active path from $\neg F$ to E . . . 129

37 Realistic switch . . . 131

60 Realistic basic switch . . . 133

61 Supply of oxygen . . . 138

62 Common cause . . . 165

63 Causal chain . . . 177

64 Freezing of water . . . 181

65 Freezing of water simplified . . . 183

66 Interdependence of causes in a common cause scenario . . . 184

67 Common cause embedded in two disjunctive scenarios . . . 186

68 Combination of a disjunctive with a conjunctive scenario . . 190

69 Collision between a and b at $\vec{x}_2$ . . . 240

70 Counterfactual scenario 1: no collision at $\vec{x}_2$ and past is fixed 241

71 Counterfactual scenario 2: no collision at $\vec{x_2}$ and future is fixed241

72 Causal graph and causal model of the bilking argument . . . 245

73 Backward causation in a scenario of quantum entanglement 247

74 Fork with a common cause 252

75 Causal chain 253

76 Combination of an open fork with a closed fork 254

77 Combination of an open fork with a closed fork, where a direction is assigned to the edges in the open fork. 255

78 Combination of an open fork with a closed fork, where a direction is assigned to all edges. 259

79 Overdetermination of the past 262

80 Overdetermination of an effect E 262

81 Two forks connected by single directed edge 266

82 Exponential growth of concurrent events 270

83 Overlap of two conjunctive scenarios 272

69 Collision between a and b at $\vec{x_2}$ 275

84 Backward causation in a scenario of quantum entanglement 280

85 Backward causation: the fork XMS is open to the past. . . . 281

86 Causal graph of the rock-throwing scenario 311

Chapter 1

Introduction

Causation is familiar but mysterious. We understand that a cause brings about its effect. The sunlight generated heat on the surface. An electrostatic discharge caused the thunder. His quick reaction saved the child from drowning. We understand these causal claims swiftly and know how to use them. Causes help us explain what is going on around us. Causes help us intervene in the course of events to bring about certain desirable effects, or prevent undesirable ones from occurring. Without a concept of causation, we could not assign responsibility for one's actions.

The mystery arises when we try to analyse causation. The relation between cause and effect seems to resist a clean analysis. Up to now, any philosophical account of causation is plagued by counterexamples. And the accounts that tally best with our common sense of what causes what usually assume primitive causal relations, and so give up on a fully reductive analysis of causation. In the light of the pervasiveness, familiarity, and importance of causation, it is astonishing that no philosophical analysis has yet succeeded. And so the challenge to find a unified theory of causation continues to be intriguing.

The aim of this book is to analyse our concept of causation. This endeavour is not so much different from devising a scientific theory: our causal judgements are the phenomena for which we seek a unified account. A primary objective is therefore to achieve an extensionally adequate theory of our causal judgements. To be precise, we analyse causal judgements in scenarios which are described in terms of deterministic laws and relations, broadly construed. The probabilistic concept of causation is not a subject matter of this book.

Another objective is to respect an empiricist principle which can be traced back to Hume's *Treatise of Human Nature* (1739/2001). A proper analysis of causation may not take any causal or modal notions as primitively given. We aim to account for our ability to make causal judgements in terms of concepts which are less mysterious and less theoretical than causation is. With some qualifications, our final theory of causation will be a reductive one.

The endeavour to analyse causation seems naive in light of past failures and challenges which have remained unanswered by even the most sophisticated contemporary accounts. Some fundamental change in approach seems necessary to come closer to a unified and extensionally adequate theory of causation. In what follows, we outline our epochetic approach to causation and explain why it can do better than extant accounts.

1 The Epochetic Approach

We aim to reconstruct how a candidate cause brought about a given effect. To this end, we begin with an analysis of what it is for a proposition to be a reason for another proposition. We explain the notion of reason by what we call an *epochetic conditional*: $A \gg B$ iff (if and only if) after suspending judgment about A and B, we can infer B from the supposition of A in the context of other propositions which we continue to believe after the suspension of judgement.

We call this approach to conditionals *epochetic* in honour of Husserl's phenomenology. Husserl (1913/1989) recommended to begin the phenomenological analysis with an operation of bracketing any judgement as to whether an object of experience has independent existence. He called this operation *epoché*, a term which is adopted from the Pyrrhonian sceptics. The goal is to capture the object exactly as it is experienced by the subject of the phenomenological analysis. Our operation of suspending judgement is motivated by a related, but different goal: to examine inferential relations between the antecedent and consequent of a conditional so as to capture some relation of dependence between the two.

The epochetic approach to conditionals suggests an epochetic approach to the notion of reason. We say that A is a reason for B iff A and B are believed, and $A \gg B$. Since we are interested in an analysis of our causal judgements, the epochetic conditional $\gg$ is relativized to an epistemic state of an agent. It is a preliminary and simple inferential analysis of the notion of reason, which however proves sufficiently powerful to build an analysis of causation on top of it.

How do we get from reasons to causes? Further constraints on the inferential relations between the antecedent and consequent of our epochetic conditional are needed for this. Throughout this investigation, we study inferential pathways from causes to effects. Novel ways to distinguish between genuine and non-genuine causes will emerge from this study. Some inferential pathways turn out to be distinctive of genuine causal relations. Others do not.

Our main thesis is that the study of inferential pathways—from causes to their effects—allows for a more accurate account of causation than extant counterfactual and regularity theories. It does seem to matter how we can infer the effect from a candidate cause. Mere inferability and regularity are not enough. Likewise, mere counterfactual difference-making, however refined, does not suffice. At the core, our approach to causation is to reconstruct causal pathways in terms of inferential pathways on the basis of a prior operation of suspending judgement.

We will give an overview of the inferential constraints in terms of which our theory characterizes genuine causal relations below. Not surprisingly, a single constraint will not do. But we keep the complexity of our theory at a moderate and intuitively accessible level. The reconstruction of causal pathways in terms of inferential ones is motivated by ideas and intuitions about causation as production. Hall's seminal paper on the distinction between two concepts of causation—counterfactual dependence and production—served as an important source of inspiration (Hall 2004). We aim to show that our causal concept of production leads to a more accurate and unified theory of our causal judgements.

How do we know which candidate causes are genuine and which are not? We simply take commonsensical causal judgements—to the extent they are widely agreed on—as phenomena to be captured by a unified theory of causation. There is a surprisingly high degree of consensus as to which events are considered genuine causes of corresponding effects. It is uncontroversial, for example, that preempted causes do not qualify as genuine. Most people agree that in a scenario of overdetermination all overdetermining causes are genuine. While such intuitions may be overruled by theoretical considerations, it has remained a desideratum to capture our commonsensical causal judgement to a maximally possible extent. We will further explain the overall methodology of our investigation in the next section.

Progress in philosophical research on causation was often enabled by developments in logic. Mackie's INUS account is based on the logic of Boolean connectives, which was not available to Hume. Lewis's counterfactual account is based on the theory of variably strict conditionals, which was developed together with the latter. Causal models in terms of structural equations have given rise to novel counterfactual accounts. In a similar vein, we exploit further resources of modern logic to analyse our non-probabilistic concept of causation. Specifically, we take advantage of belief revision theory to define an epochetic operation of suspending judgement. Logical accounts of default and abductive reasoning will be used to study inferential and causal pathways. While the book is technical at times, it is self-contained and requires only familiarity with classical logic. We have worked hard to make it as accessible as possible to all readers.

2 The Concept of Causation

In his *Treatise*, Hume first considers the idea that causation is some necessary connection between cause and effect. His criticism is strong. Such a connection cannot be observed in concrete instances of causation. Nor can a necessary connection between cause and effect be demonstrated by reason alone. The attempt to give such a demonstration leads to the problem of inductive reasoning.

Hume moves on to suggest an alternative: causation may be understood in terms of a regular connection between cause and effect. This account, however, may also be criticized with reference to the problem of inductive reasoning: which reasons do we have to think that a certain regularity observed in the past will also hold in the future? Hume addressed this problem by explaining the origin of our idea of regular connection in terms of custom and habit. This has led him to an epistemic interpretation of the concept of regular connection in his analysis of causation:

> A cause is an object precedent and contiguous to another, and so united with it, that the idea of the one *determines the mind to form the idea of the other*, and the impression of the one to form a more lively idea of the other. (Hume 1739/2001, Book I, Part 3, Sect. 14, § 31, our emphasis)

Hume even went on to suggest an epistemic account of necessity in order to make sense of the old idea of necessary connection between cause and effect (Hume 1739/2001, Sect. 14 in Part III of Book I). Thus we find epistemic and non-epistemic accounts of the condition of regular connection in Hume's *Treatise*. Why should the epistemic account fare better with regard to the problem of inductive reasoning? Arguably, some determination of the mind is accessible at the time of the respective causal judgement. We can be aware of such a determination without knowing whether or not the respective regular connection continues to hold in the future.

We must nonetheless wonder how the epistemic account of regular connection is related to the non-epistemic one. Is one more fundamental than the other when it comes to understanding what causation is? This question has troubled Hume scholars for a long time (see, e.g., Robinson (1962), Richards (1965), and Beebee (2011)).

More contemporary work on causation is facing a similar tension. Mackie (1980, p. 1), for example, understands the analysis of causation as an ontological project. It must answer the question of what causation is *in the objects*. This qualification is deemed misleading by Hausman (1998, p. 8n). He argues that it is rather futile to distinguish between an inquiry into the meaning of the causal concept and an analysis of causation in the objects. In defence of Mackie's methodology, one might reply that no such distinction was envisioned. Paul and Hall (2013, p. 249) hold that a unified theory of causation may have the form of an ontological reduction or a conceptual analysis. We follow suit. To our mind, it may well be fruitful to pursue the two projects independently from one another, and explore their relationship at a later stage of inquiry.

Our theory of causation clearly has the form of a conceptual analysis. It qualifies as such in at least two dimensions. First, we take our causal judgements in science and everyday life as phenomena for which we seek a unified account. Second, we aim to tell a story of how human minds come to make such judgements which is cognitively plausible. We understand the notion of cognitive plausibility along broadly empiricist lines. Basically, our empiricism comes down to the belief that some concepts are less mysterious than others. We think that a concept A is less mysterious than a concept B iff A can be used and applied without relying on applications of concept B, at least in the majority of its applications. For example, we believe that spatial relations among events—which are observable by unaided perception—are less mysterious than, say, the concept of electromagnetic force.

To give an example, suppose we see an apple falling toward the ground. We think this motion is caused by the gravitational force of the Earth. But the motion of falling can be observed independently of the causal relation and the gravitational force. It seems impossible, by contrast, to observe gravitational forces independently of the trajectories of concrete objects. This conceptual order seems invariant under any training.

To assume that some concepts are less mysterious than others is not to assume that there are concepts which are free of any mystery. We merely believe that some concepts are more theoretical—and in this sense more mysterious—than others. Our empiricism is inspired by work in philosophy of science which may be described as *post-logical empiricist* in that a number of doctrines of early logical empiricism are dropped, without however giving up entirely the project of a rational reconstruction of science. Specifically, people in the structuralist school have tried to recognize an ordering of theoreticity among the concepts which are used in the applications of scientific theories. This order may well be a partial one, specifically if networks of scientific theories are considered.[1]

When setting out to analyse causation, Hume took it for granted that temporal relations among observable events are less mysterious than causal relations. In Part II of this book, we aim to show that Hume was on the right track and that contemporary alternatives to the Humean convention fail to give us a comprehensive account. Our reductive theory of causation is motivated by the belief that temporal relations are less mysterious than

[1] See Sneed (1979) and Balzer et al. (1987) for the canonical expositions of the structuralist approach to science. See Andreas (2020, Ch. 5) for an axiomatic account of the structuralist representation scheme and some further work on the structuralist criterion of theoreticity.

causation, but it does not depend on this belief. From a logical point of view, it makes perfect sense to analyse a concept A in terms of concepts B, C, and D, without assuming any conceptual hierarchy. The presumed order of theoreticity and mysticity remains nonetheless important for the theory to be read as an epistemological story about causation.

Our epistemological story makes some idealizing assumptions about the logical capacities of the human mind. Most notably, the analysis of the inferential pathways from candidate causes to their effects is spelled out in terms of natural deduction. This inferential approach is motivated by ideas about causation as production. But we do not claim that our brain is at bottom a logical machinery. The use of logical concepts is a methodological principle, comparable to the use of mathematics in science.

An important motivation for us to once again take up the Humean project of analysing causation is that we have nowadays much more advanced and refined logical tools available than Hume and even Carnap had at their time. Specifically, we will use concepts of belief revision theory for our epistemological story about causation. The use of logical concepts will be justified by the result of our investigation—a unified theory of causation. Further justifications may be given, but we will not discuss the relationship between logic and philosophy here.

Ideally, an epistemological analysis of causation should be in harmony with some viable metaphysics of causation in the objects. The former analysis should make plausible how our causal judgements capture causal relations in the world. Reversely, a viable metaphysics of causation should be connected with some epistemological story of how human minds come to make causal judgements.

We will outline in the Conclusion how our epistemic analysis may be extended to a theory of causation in the objects. The focus of this book is nonetheless on the epistemology of causation by way of a rational reconstruction of our concept of causation. Our theory is aimed to be as metaphysically neutral as possible in that it may be connected to a variety of different views about human minds and the world. For this reason, we will avoid as much as possible commitments to specific accounts of propositional meaning, events and absences, event types, etc. Further connections to the metaphysics of causation may be explored at a later stage of inquiry.

3 Reasons

Our account of reasons originates from an epistemic approach to causation due to Ramsey. As is well known, Ramsey (1931a, footnote 1) proposed the following evaluation procedure for conditionals which is known as the Ramsey Test:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; so that in a sense 'If p, q' and 'If p, $\overline{q}$' are contradictories.

This epistemic evaluation recipe for conditionals has received probabilistic and non-probabilistic interpretations.[2] The core of the latter interpretation has been pointedly expressed by Stalnaker (1968, p. 102):

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

Now, we suggest a subtle variant to this test. Instead of adding the antecedent right away and then making adjustments to our beliefs in order to ensure consistency, we begin with suspending judgement on the antecedent and the consequent. The subsequent operation of adding the antecedent will then always lead to a consistent set of beliefs. In brief, we suggest to strengthen the Ramsey Test as follows:

> First, suspend judgement about the antecedent and the consequent. Second, add the antecedent (hypothetically) to your stock of explicit beliefs. Finally, consider whether or not the consequent can be inferred from your explicit beliefs.

[2] For a unified account of probabilistic and non-probabilistic interpretations of the Ramsey Test, see Günther and Sisti (2022).

Let us write $A \gg B$ iff B can be inferred from A after judgement has been suspended about A and B. Then we say that A is a reason for B iff A and B are believed, and $A \gg B$. This notion of reason is obviously relative to the beliefs of an epistemic agent. After all, we aim to devise a theory of our causal judgements.[3]

This account of the notion of reason is preliminary since it faces some obvious counterexamples. Specifically, there remain symmetry problems. For example, given A, B, and $A \wedge B$ are explicit beliefs, we have to say that A is a reason for $A \wedge B$, and vice versa. The latter seems counterintuitive. Such symmetry problems, however, disappear once we impose further constraints on the inferential relations between antecedent and consequent of our epochetic conditional.

The relation of reason is commonly considered a relation between propositions rather than sentences. When writing $A \gg B$, we assume that A and B are sentences which have some form of propositional meaning. We do not explore or offer any theory of propositions here. The reader may think of propositions in terms of whatever account she thinks meets best her philosophical desiderata. We assume only that the account of propositions is consistent with the principles of classical logic.

4 Causes in Causal Models

How do we get from an analysis of the notion of reason to a theory of causation? How do we get from reasons to causes? The key idea is to reconstruct how a given cause may have brought about its effect along an active path. The notion of active path is explained in terms of inferential pathways. We will now give an overview of the conditions in terms of which we characterize reasons which stand for genuine causes.

Our analysis is divided into two parts. In Part I, our task is a little easier since we use causal models along the lines of Halpern (2000) and Pearl (2009). Such models take certain causal relations as elementary and primitive. A causal model analysis is not reductive unless the notion of structural equation is explained in non-causal terms. A great deal of research has been

[3] We assume here that the antecedent of $\gg$ is not a contradiction and the consequent is not a tautology. For a technical way to restrict conditionals to contingent antecedents, see Günther (2022).

done on how to define causation relative to a causal model. But it proved surprisingly difficult to analyse complex causal relations, even if the elementary ones are given by a causal model. The challenge is to define the notion of cause in such a manner that our causal judgements are captured as comprehensively as possible in a large variety of different scenarios. For this to be achieved, causal scenarios known as *overdetermination*, *early and late preemption*, *prevention*, *omission*, *switches* as well as a number of variants and combinations thereof have been studied extensively.

To make further progress, we devise a system of natural deduction with structural equations. This system basically complements the semantic account of causal models in Halpern (2000). Then we define our epochetic conditional $\gg$ for causal models in terms of natural deduction. Once such a conditional is in place, only two inferential constraints are needed to capture genuine causal relations.

Suppose it holds that $C \gg E$, where the conditional $\gg$ is defined for causal models. This means that there is a deduction of E from the assumption C, which uses further premises in the form of structural equations and, possibly, information about the context of the presumed causal process. The set of further premises neither implies C nor E since we suspended judgement on the candidate cause and its effect. When is C a genuine cause of E?

First, we require that there is an active path from C to E. An active path from C to E is an inferential path such that each inferential step—on the way from C to E—depends on the assumption of C as a premise in the deduction of E. The rationale for this is that each inferential step to an event or absence may be interpreted as a section in a causal process which was started by the candidate cause. We will show that preempted causes violate the condition. If you like, we reconstruct a concept of factual dependence of the effect on the candidate cause.

Second, genuine causes are required to satisfy a condition of deviancy. A genuine cause must be at least weakly deviant—in the sense that it goes against a norm or default law. Also, if there are events or absences in the context of the candidate cause which are weakly normal, then we must not suspend judgement on them when looking for an agnostic model with an active path. We call an event or absence weakly normal iff it conforms to a norm or default.

Such are the cornerstones of our analysis in Part I. In sum, causation is inferential dependence along causal pathways such that each section of each pathway depends on the candidate cause. We will explain the notion of inferential dependence in terms of an epochetic conditional and natural deduction. This notion is relative to an epistemic state. Also, genuine causes and their context satisfy a condition of deviancy and normality, respectively. Consideration of deviancy is needed to analyse causation by omission, some scenarios of prevention, and realistic switches. The deviancy condition applies nonetheless to genuine causes in other scenarios as well since our account of deviancy implies that any occurring event is at least weakly deviant.

The primary objective of Part I is to analyse causation such that our judgements in concrete causal scenarios are captured as adequately as possible. Our analysis makes important progress on this project, sometimes referred to as the study of *actual causation*. It is the first analysis to capture our commonsensical causal judgements in all of the scenarios studied throughout chapters 3 to 6 and beyond. It captures, in particular, our causal judgements in scenarios such as overdetermination, preemption, short circuits, switches, prevention, omissions, as well as variants and combinations thereof. The currently popular counterfactual accounts already face serious problems with the just mentioned scenarios.

We argue elsewhere in detail that our epochetic analysis outcompetes the most advanced counterfactual accounts with respect to the adequacy of our causal judgements (Andreas and Günther 2025a). While we think we have made substantive progress in analysing causation, we don't think that we have solved all problems of actual causation. This is why we would like to encourage our readers to challenge and to improve the analysis presented here.[4]

[4] The analysis of this book evolved out of a series of epochetic analyses, published in Andreas and Günther (2021a,b, forthcominga) and (forthcomingb). It goes beyond the latter work in particular by the study of causal pathways via the notion of an active path. Otherwise, the analysis presented in Andreas and Günther (forthcomingb) is almost identical to the one in Part I. It's finally worth noting that our epochetic analysis gives rise to a regularity theory of actual causation (Andreas and Günther 2024a,b).

5 The Reductive Analysis

In Part II we take up the challenge to devise a reductive analysis. No structural equations are available anymore. We begin with an explanation of the basic concepts of belief revision theory in the tradition of Gärdenfors (1988). Then we define an epochetic conditional for sets of explicit beliefs.

Again, suppose it holds that $C \gg E$, where the conditional $\gg$ is now defined for a language of classical propositional logic or classical first-order logic. This means that there is a deduction of E from the assumption C in an epistemic state which is uninformative on C and E. The deduction uses laws of a background theory as premises. Importantly, we work only with a minimalist, non-modal, and syntactic notion of law, which is weaker and less demanding than some substantial notion of a law of nature. When is C a genuine cause of E?

At this point, the Humean convention comes into play: a cause precedes its effect. Of course, we cannot adopt this convention in a simple and straightforward manner. Specifically, we need to address well-known problems and objections to this convention: the problem of spurious causation, instances of simultaneous causation, and the conceptual possibility of backward causation. Part II is centred on these problems.

The problem of spurious causation is perhaps the hardest to be dealt with when one wishes to reconsider a Humean approach to causation. The problem may be summarized by the well-known dictum that correlation is not causation. Sometimes there is a correlation between two events such that one precedes the other, but we do not consider the preceding event a cause of the other. The correlation between a drop of the barometer and stormy weather is a famous example, if only a probabilistic one. For a deterministic example, consider an event of lightning. A bright lightning flash in the sky is followed by thunder. The correlation is relatively strict, depending on further details of the description of the events. But we do not consider the flash a cause of the thunder. The correlation is rather due to a common cause, given by an electrostatic discharge between a cloud and the ground.

We solve the problem of spurious causation by two constraints on the inferential pathway from the candidate cause C to the effect E. First, it is required that each inferential step to an event or absence is forward-directed in time in the weak sense of not being backward-directed. No event or absence asserted in the premises of an inferential step—to an event or absence—may be temporally later than the occurrence or absence inferred. This condition may be described as a proof-theoretic variant of the Humean convention. Roughly, the Humean convention must not only hold for the pair of cause and effect, but also for each inferential step from the candidate cause to the effect.

Second, we require that each law used on the inferential path from *C* to *E* is non-redundant. The notion of law itself is minimalist, non-modal, and syntactic, as indicated above. Once such a minimalist notion of law is in place, we impose the constraint of non-redundancy. We define this constraint by means of a formal system of abductive reasoning. In the background is the best system account of laws in philosophy of science. Once again we draw on the work of Ramsey, who suggested such an account of laws in his essay on causation (1931a, p. 242).

The first formulation of the best system account goes back to Mill (1843/2011, Book III, Ch. IV). Notably, Mill also suggested to overcome problems of spurious causation by invoking some notion of nomic regularity. The latter notion in turn is explained by his best system account. Our contribution to this line of research is to define the notion of non-redundant law in terms of an inference system of abductive reasoning and an inferential variant of the Humean convention. No causal notions are used in our account of abductive reasoning.

We show that our analysis avoids problems of spurious causation for a number of well-known scenarios. Specifically, we show that it delivers the intuitive verdicts for all causal scenarios in which the common cause is embedded in a conjunctive or a disjunctive scenario. Moreover, we apply our analysis to scenarios of spurious causation for which such an embedding is less easy to identify. Our solution to the problem of spurious causation has some interesting connections to the regularity theories by May and Graßhoff (2001), and Baumgartner (2013), and the independence theory by Hausman (1998).

The problem of simultaneous causation turns out to be less challenging. We study some scenarios of this type of causation, and come to the following observation: there is an asymmetry between the simultaneous cause and its effect in that we have a Humean causal explanation for the cause which is independent of the effect, but not vice versa. A Humean causal explanation is simply one which satisfies the Humean convention. Thus we can account for simultaneous causal relations in terms of Humean causal explanations.

There remains to discuss the problem of backward causation. Obviously, if there is backward causation in our world, the Humean convention is not tenable, at least not without further modification. The latter qualification is important, though. If we had some criterion of backward causation in place, the Humean convention may still be used as default criterion for the distinction between causes and their effects. Dowe's (1996) account of backward causation is guided by a similar line of reasoning. He proposed a disjunctive approach to the direction of causation, which is based on two different, albeit related criteria to distinguish between causes and effects. One of the two criteria is, at least extensionally, equivalent with the Humean convention. We outline how Dowe's proposal may be adopted for our reductive analysis of causation.

At the same time, we come to observe that there is no commonly agreed understanding of the notion of backward causation to be found in the literature. It has remained very much an open problem what it could mean to find evidence for backward-directed causal relations in the world. We show that Price's (1996) very nuanced and well motivated account of backward causation amounts to a disjunctive approach to the direction of causation as well. Surprisingly enough, one of the two disjuncts is based quite directly on the Humean convention. This result indicates that a disjunctive approach may be unavoidable when one wishes to allow for the possibility of backward causation. The problem is that we are lacking a unified account of forward and backward causation which explains to us what kind of evidence, however indirect, we could in principle obtain for the two types of causation, respectively. Lewis's (1973a) counterfactual analysis of causation is no viable option for such a unified account.

Once again the primary objective of our analysis in Part II is to develop an extensionally adequate theory of causation—adequate with respect to our causal judgements in science and everyday contexts. This theory is aimed to be reductive in that it does not take any causal or modal notions as primitively given. Nor do we take some distinction between laws of nature and accidental generalizations for granted.

Our reductive theory may be summarized as follows. Causation is nomic inferability of the effect from the cause—in an epistemic state which is uninformative on the cause and the effect. No inference to an event or absence on the inferential path from the cause to the effect must be backward-directed in time. Causes precede their effects—unless the cause is simultaneous with the effect, and we have a causal Humean explanation of the cause which is independent of the effect, but not vice versa.

Finally, we will show in the Conclusion how our reductive analysis may be used as a foundation of deterministic causal models with structural equations, thus establishing a connection between the causal model analysis in Part I and the reductive analysis in Part II. We do not offer a translation of non-causal laws into structural equations, but explain what it could mean that a causal model with structural equations is verified by an epistemic state which does not contain explicitly causal beliefs, given our reductive theory. Causal models are thus traced back to epistemic states with beliefs about causal scenarios, where these beliefs are couched in a language free of causal and modal notions.

Spohn's (2006, 2012) work on causation has served as an important source of inspiration for Part II. Most notably, Spohn begins with an analysis of the notion of reason, and then builds an analysis of causation on top of it. Also, he makes use of the Humean convention to account for the direction of causation. At the same time, we go beyond his ranking-theoretic analysis by a detailed study of inferential and causal pathways. Specifically, we devise a properly reductive solution to the problem of spurious causation. Spohn's solution to this problem doesn't seem to meet this standard, as will be shown in Section 13 of Chapter 8. We have aimed to give the most elaborate defence of the Humean convention to be found in the literature so far.

6 Events

Causation is a relation among events. This dictum has been expressed in much of the more contemporary work on causation. We follow suit. We do not see strong enough reasons to consider alternative relata of the causal relation.

Events may or may not occur. If an event does not occur, we say that it is absent. The absence of an event is simply called an *absence*. We do not have an elaborate theory of absences, let alone negative events. Unlike a fact, it is conceptually possible for an event to be absent from the actual world.

Absences are sometimes considered causes of events, and vice versa. We should therefore clarify that causation is a relation among events and absences. Causation by an absence is troublesome for the metaphysics of causation. But it is less so for an analysis of the conceptual relations between causal and non-causal statements.

Events are commonly thought to be concrete. An event concerns a particular object in space and time, a relation among such objects, or a specific spatiotemporal region. Again, we follow suit. It is worth noting, though, that events are often not fully specified down to the level of particular objects in the discussion of a causal scenario.

Take the famous example of two children throwing rocks at a bottle. Suzy's rock hits first, and causes the bottle to shatter. Billy's rock comes a little later, and so doesn't get to hit the bottle. His throw of a rock is preempted. Even though we know the names of the two children, the events are not specified to the extent that we can identify concrete events with a precise spatiotemporal location. Suzy and Billy are more like fictional characters in a causal story, which is supposed to deliver a more general insight into a class of causal scenarios.

The lack of specificity is even more obvious in the case of neuron diagrams. Such diagrams are so simple that it seems impossible to identify uniquely a concrete neuronal system which a given diagram is supposed to represent. A neuron diagram in a philosophy paper seems rather a simplified and schematic representation of a class of causal scenarios.

We must wonder whether the lack of specificity of a causal model is a virtue or a problem. Suppose the design and the study of a causal model is confined to token-level causation. This is the level of concrete and specific events. Then the lack of specificity is certainly a problem. It may be addressed by, charitably, assuming that some details are left out from the causal model for simplicity. The idea is that, ultimately, the author of a causal model always has a concrete scenario of causation in mind. This way of fixing the problem may or may not be satisfactory, but let's assume it is.

Suppose now, by contrast, the variables of a causal model are intentionally not fully specified. They stand for event types rather than specific events.

The instances of the event types must satisfy certain constraints. For example, one throw of a rock may stand in the relation of preceding another with respect to a target. But these constraints do not enable us to identify specific events. They are like the axioms of a scientific theory which have different interpretations, all of which satisfy the axioms equally well. The Peano axioms are a case in point. The successor function and the symbol for the number zero may be interpreted in different ways without violating any of these axioms. In brief, any infinite sequence of objects satisfies the Peano axioms. In a similar vein, we can interpret the variables for Suzy's and Billy's throw of a rock by different, albeit related events without violating any of the structural equations of the causal model.

On the latter perspective, a causal model determines a complex concept which may be applied to a variety of different, fully specified causal scenarios. Token-level causation comes into play when a causal model is applied to a scenario of fully specified events. In the absence of such an application, all causal claims concerning the model are at the type level. While causal models of scenarios such as overdetermination and preemption may be less general than the models devised by scientists, they exhibit nonetheless some level of generality. Causal modelling is arguably not so much different from applying some axioms of whatever framework of scientific theories to certain phenomena. The view that a causal model defines a complex concept—which applies to a range of concrete scenarios—may be further elaborated by means of set-theoretic predicates, as explained in Suppes (1957) and Sneed (1979).

Our analyses of causation are open to both readings of causal models. We find the more abstract view attractive, but it is not an assumption of our theory of causation. If adopted, the variables of a causal model stand for event types or, equivalently, incompletely specified events. If rejected, the variables are to be read as referring to specific events. To distinguish between specific events and event types is not to deny that events are concrete. Event types may be conceived of as classifications of events.

Let's briefly exemplify our claim that a causal model may be used to capture causal relations at the type-level. In a deterministic setting, it is rare to find cases in which the presence of one property causes the presence of another without further qualifications. Standard examples of type-level causal claims are taken from probabilistic causation, such as the statement that smoking causes cancer. However, we may have simple deterministic causal relations at the type-level in at least some idealized systems of physics. Take the claim that total Newtonian forces cause a body to accelerate. That's universally true in all systems described by classical mechanics.

We can represent the causal relation in question by a causal model in a straightforward manner. Let the variable F stand for the presence of a total Newtonian force on an object a. Further, let A stand for the object a having an acceleration whose value is greater than zero. Using the standard notation of causal models from Pearl (2009), we can simply write $A = F$ to express a very simple deterministic causal relation. Note that the letter a merely serves as placeholder for an object which is not further specified. This is how we can express generality of the causal model.

Things are more complex when we study causal models of overdetermination, preemption, etc. Then we study causal relations in types of causal scenarios rather than type-level causal relations which are supposed to hold between properties directly. The activation of a neuron in a scenario of overdetermination, for example, may be understood as a placeholder for events which have in common that another event is present which is equally sufficient for some effect in question. On this reading, causal models specify types of causal scenarios.

For clarification, it may be helpful to distinguish between two types of type-level causal relations. One concerns relations among properties, as in the statement that total forces cause bodies to accelerate. Another is about causal relations in types of causal scenarios. The former type is of little interest in the context of deterministic causation. But the latter type has attracted a great deal of research for quite some time.

Furthermore, a note on the relationship between events and their description is in order. The description of an event may be fine-grained or coarse-grained. We can speak, for example, of the event that the tennis ball hit the ground behind the baseline. Or we can say that the tennis ball hit the ground exactly two metres behind the baseline. Also, we can say that the tennis ball hit the ground exactly two metres behind the baseline at a specific location and a specific time, say the Northern side of the tennis court at Roland Garros Stadium in Paris on 8 June in 2025 at 4:05:30 pm.

The different descriptions in question may be relevant for the evaluation of causal claims. Suppose we are indeed interested in the causes of a concrete event of a tennis ball hitting the ground behind the baseline at your local tennis court. Further, suppose the wind was moderate, but it was not completely calm. Then it makes a difference whether or not we furthermore specify that the ball hit the ground exactly two metres behind the baseline. This is so because it's entirely plausible to say that the wind is a cause of the ball hitting the ground two metres behind the baseline, if only a conjunctive one. It's one causal factor of the effect in question. By contrast, we should not say that the wind caused the tennis ball to go out behind the baseline because moderate wind never drives a tennis ball away by two metres.

The relationship between events and their description is an open philosophical problem. Do phrases like 'Caesar's death', 'Brutus's killing Caesar', and 'Brutus's stabbing Caesar' refer to the same event? Kim (1969) pointed out substitution problems in explanatory and causal statements if we affirm this question. We have just observed an instance of such a problem: replacing a coarse-grained description of an event by a more fine-grained one in a causal statement may well lead to a change of truth value. Davidson (1969) nonetheless maintained that events may be redescribed in wildly different ways. The above phrases concerning the death of Caesar all refer to the same event. Despite apparent substitution problems, Davidson held on to the view that events are the relata of the causal relation.

The nature of events and their identity criteria is a pressing problem for a fully-fledged metaphysical account of causation. It is less pressing for our endeavour to analyse the concept of causation. In Part I of this investigation, our analysis is relative to a causal model. In Part II it is relative to an epistemic state. Events are therefore always given to us through a description and conceptualization. Such descriptions may be fine-grained or course-grained. There is no need to make a decision as to whether events, in general, are fine-grained or coarse-grained. However, we should clarify that it's events under a description which are the relata of causal relations in our analysis.

There remains to discuss one problem concerning events. Our frameworks in both Part I and II use variables and sentences which stand for events. Do these variables and sentences have to stand for distinct events? The requirement in question is crucial to reductive counterfactual approaches in the tradition of Lewis (1973a). Suppose we have a variable for precipitation and one for snowfall. Then it holds true that, had there been no precipitation, there would have been no snowfall. However, we should not infer from this that precipitation is a cause of snowfall.

Thanks to the Humean convention, our analysis avoids mistaken causal verdicts arising from non-distinct variables, at least to a large extent. Since there is no temporal difference between snowfall and precipitation in the scenario in question, neither counts as cause of the other. Nor is the tennis ball hitting the ground two metres after the baseline a cause of the tennis ball going out. Fine-grained and coarse-grained descriptions of what we intuitively consider to be the same event may well co-exist in one and the same account of a causal scenario.

Things are less clear-cut when we study scenarios of simultaneous causation in Part II, though. In the absence of an explicit requirement that the variables of a model refer to distinct events, some relations of grounding qualify as relations of simultaneous causation. This may either be a problem or a virtue, depending on your view of the relationship between grounding and causation. Brentano, for example, considered causation as a type of grounding.[5] Wilson (2018) has argued for the view that grounding is a type of causation. We will take a closer look at some scenarios of grounding and simultaneous causation in Chapter 9, but do not attempt at a general characterization of grounding here. Let's therefore take the requirement that the variables of a model refer to distinct events as optional for our analysis.

The subject matter of this book is deterministic causation. Broadly construed, deterministic causation is not confined to causal scenarios which are governed by strict laws which hold without any exception. This is not so for two reasons. First, some exceptions to the laws may be dealt with by acknowledging that a causal model almost always makes idealizing assumptions. Causal models are just as imperfect as a number of scientific models are. Second, we have nowadays formal systems of reasoning which account for the use of ceteris paribus laws without probabilistic notions. Details will be explained in Part II.

One word on pluralism about causation is in order. This position has been suggested and recommended along different dimensions. Most obviously, we distinguish between probabilistic and deterministic causation. Moreover, the distinction between type and token-level causation has received a great deal of attention in the literature. Hall (2004) suggested to distinguish between causation as production and counterfactual dependence. Hitch-

[5] See Schnieder (2014) for a detailed discussion.

cock (2007) critically examines the distinction between the scientific, the folk attributive, and the metaphysical concept of causation. More recently, Fischer (2024) argued for the distinction between total, path-changing, and contributing actual causation. This list of distinctions is not exhaustive.

Our theory aims to be as monist as possible about deterministic causation, and so we don't adopt a pluralist position. We have just argued that it captures deterministic causal relations at both the token and the type level. Part I is concerned with everyday causal judgements, while Part II focuses on causal relations in science. The two parts are merged into a unified theory in the final chapter of the book. For now, the theory is confined to deterministic causation, broadly construed. However, there is work underway which generalizes our theory to probabilistic causal relations.

7 Advantages over the Counterfactual Approach

Our epochetic theory says that causation is inferential dependence along an active path from cause to effect. Through this notion of inferential dependence we aim to reconstruct the idea that a given effect factually depends on its cause. In brief, inferential dependence reconstructs some concept of factual dependence. No notion of counterfactual dependence is needed in our theory. In this sense, it is a proper alternative to the currently popular counterfactual accounts of causation.

Our theory makes important progress. It has, in particular, three major advantages over the most advanced counterfactual accounts. First, it captures more causal scenarios and some for the first time. Second, it is conceptually more unified than the most advanced counterfactual accounts. Third, our theory is properly reductive. Let's briefly explain why our theory has these advantages.

In Part I, we develop our causal model analysis of causation and show that it captures our commonsensical causal judgements in all of the scenarios studied in chapters 3 to 6. No other account we are aware of—counterfactual or not—achieves this result. Our analysis is the first to account for our causal intuitions concerning some switches. It furthermore succeeds in capturing the causal nature of both norm-compliant actions and norm-deviant omissions. More generally, it recognizes normal events and deviant absences as causes without falling prey to a proliferation of causes.

Another important result concerns the problem of Schaffer's (2000) trumping preemption. We solve this problem using only a minimal model of the causal scenario. For this result to be achieved, it is essential that our framework of causal models comes with both a model-theoretic semantics and a system of natural deduction. The latter is exploited in our notion of an active path from cause to effect. By contrast, the causal model accounts due to Hitchcock (2001), Halpern and Pearl (2005), Halpern (2015), and Gallow (2021) rely on a purely semantic account of structural equations. They cannot distinguish the genuine cause from the trumped cause in the minimal model since, for them, the minimal model of trumping is semantically indistinguishable from the causal model of the overdetermination scenario. All the extant causal model accounts using Pearl's (2009) framework of causal models face a problem here which derives from the foundation of their framework. Further problems of the counterfactual approach to capture our causal judgements are detailed in Andreas and Günther (2025a).

The most advanced counterfactual accounts face conceptual tensions in their attempts to capture our causal judgements. Halpern and Pearl's (2005) account of causation aims to reconstruct an *active causal process*, and Gallow's (2021) the *transmission of deviancy in an uninterrupted process*. However, Halpern and Pearl's account is based on two disparate ideas, as they acknowledge with reference to Hall's 'Two Concepts of Causation' (2004):

> Our definition certainly has some features of both counterfactual dependence and of production—AC2(a) captures some of the intuition of counterfactual dependence (. . .) and AC2(b) captures some of the features of production (. . .). (p. 867)

Our epochetic theory understands causation as a type of production along an active path. No notion of counterfactual dependence is needed in addition. And so our theory faces no conceptual tension between two disparate ideas of causation. In this sense, our theory is conceptually more unified than Halpern and Pearl's (2005).

Gallow's (2021) theory of causation looks conceptually well unified. On closer inspection, however, his theory deviates from its own motivating idea, which leads to conceptual tensions. For this to be seen, consider his outline:

> the theory says that *C* caused *E* whenever both *C* and *E* are deviant or non-inertial events, and there is an uninterrupted process which *transmits* *C*'s deviancy to *E*. (p. 47)

This sounds as if the *uninterrupted* process transmits deviancy on each step of counterfactual dependence from one event or absence to the next. However, Gallow's formal implementation in terms of *causal networks* does not demand such a gapless transmission of deviancy. He deviates from his motivating idea because otherwise his theory could not recognize double preventers as causes.

Gallow's theory requires that both cause and effect are deviant. This means that normal or default events can neither be causes nor effects. As a consequence, his theory does not recognize norm-compliant actions and events as causes. There is also no causation by simple prevention. Her drinking a glass of water prevented her dehydration. Not being dehydrated is a normal absence on Gallow's theory and so cannot be caused. We see no easy way to repair this without even further deviating from the motivating idea that causation is transmission of deviancy in an uninterrupted process. Our epochetic theory faces no such problems while staying true to its core idea.

Part II presents our reductive analysis of causation. At its heart is the Humean convention: a cause precedes its effect in time. We develop rigorous solutions to the problems of spurious and simultaneous causation. Moreover, we consider the conceptual possibility of backward causation, and show that extant accounts of backward causation may be adopted by way of a disjunctive approach to the direction of causation. The Humean convention remains one of the two disjuncts. The disjunctive strategy has been suggested by Price (1996) and Dowe (2000, Ch. 8) as a means to make sense of backward causation.

By contrast, the counterfactual approach has not made much progress on the project of a reductive analysis since Lewis (1979). His idea to account for the direction of causation in terms of the semantics of counterfactuals and the overdetermination thesis is certainly original. This approach, however, runs into seemingly insurmountable problems (see Price (1996, Ch. 6), Elga (2001), and Frisch (2005, Ch. 7)). Most notably, at the micro level of at least some physical systems we cannot recognize asymmetric forks which display the overdetermination of the past. Relatedly, there is the problem of physical systems described by time-symmetric theories. We will explain these problems and point out an additional one in Section 8 of Chapter 10.

A viable alternative to Lewis's proposal is hard to find in the literature on the counterfactual approach. Noordhof's (2020) defence of this approach emphasizes the variety of causes, and so does not attempt at a unified account of the direction of causation. Hausman (1998), Spohn (2006), and Baumgartner and Falk (2019) have developed novel reductive accounts of causation, but none is of the counterfactual type. Woodward's (2003) interventionist account is of the counterfactual type, but not reductive. Menzies and Price (1993), and Price (1996) took up the challenge to develop a reductive analysis from an interventionist perspective, but the semantics of interventionist conditionals is not worked out in greater detail.

Part I

Causes in Causal Models

Chapter 2

Active Paths

How do we get from reasons to causes? To analyse causation, we impose further constraints on the inferential relations between antecedent and consequent of our epochetic conditional. More specifically, in Part I we use an inference system of causal models and structural equations. Such equations encode certain elementary causal dependences among events and absences, which are commonly taken as antecedently understood without further analysis.

In this chapter, we begin with an account of causal models on the basis of classical propositional logic. Using a simple semantics of truth values, we will define a relation of entailment for causal models. Then we outline an account of deductive reasoning with structural equations. Only two simple inference rules will be needed for this on top of classical propositional logic.

Once we have a logical account of causal models in place, we can study inferential pathways in such models, and reconstruct how a candidate cause brought about a given effect. The key idea is that each inferential step to a literal may be interpreted as a section in a causal process which was started by the candidate cause. We express this condition by a constraint on the inferential pathways leading from the candidate cause to the effect: any inferential step to an event or an absence must depend on the assumption of the candidate cause. This way we define the notion of *active path* from a candidate cause to its effect.

1 Causal Models

In this section, we introduce causal models and structural equations in a logical format. This implies a moderate deviation from the standard account by Halpern (2000) and Pearl (2009). We will distinguish more clearly than the standard account between the syntax and semantics of causal models. Most notably, we introduce both a relation of semantic entailment and an account of deductive reasoning with structural equations. The latter will open up new avenues for analysing causation. For we can then study in greater detail different types of inferential pathways from presumed causes to effects. Eventually, this study results in a novel way of discriminating between genuine and preempted causes.

We develop our account of causal models on the basis of classical propositional logic. An important benefit of this design choice is that causal models require little more than propositional logic. Only truth values of propositional formulas will be used to define the semantics of structural equations and a relation of semantic entailment. Our introduction to causal models aims to be as accessible as possible to philosophers and people from other disciplines interested in causation.

For the interested reader, we have written an appendix on the logic of causal models, in which further details and results are developed. Specifically, we prove soundness and completeness for our deductive system of causal reasoning there. Moreover, we show how the restriction to binary variables—inherent in the present account of propositional causal models—may be lifted by means of a fragment of sorted first-order logic.

The account of causal models in Part I remains restricted to binary variables, though. This restriction is not severe for mainly two reasons. First, virtually all of the widely discussed causal scenarios can be represented using only binary variables. Second, more importantly, the generalization of our analysis to non-binary variables is straightforward. Once syntax and semantics of structural equations with non-binary variables have been explained, our analysis works for causal models with such variables. No further modifications are needed. An important benefit of our design choice to build causal models on top of propositional logic is overall simplicity. Another is that we can study inferential pathways between causes and effects in greater detail.

A brief note on variable selection is in order. Standardly it is assumed that the variables of a causal model satisfy the principles of exhaustivity, exclusivity, proportionality, and distinctness (McDonald 2020). Since we use only causal models with binary variables which stand for events, the principles of exhaustivity, exclusivity, and proportionality are trivially satisfied. This claim is easy to verify. The discussion of distinctness is a bit more involved, we touched on this principle in the Introduction. Are the events represented by two different variables always strictly distinct from one another? Is it possible for the cause to be contained in the spatiotemporal region of the effect? We consider the principle of distinctness to be optional for our theory. It is needed if one wishes to maintain a strict separation between simultaneous causation and grounding. Details are explained in Section 3 of Chapter 9.

1.1 Structural Equations

A causal model represents a causal scenario. In a such a scenario, certain events occur, others do not, and there are dependences among these events and absences. We represent events and absences by literals, and the dependences by structural equations. A literal is an atomic sentence or the negation of such a sentence. In classical propositional logic, A and $\neg A$ are the two literals of the propositional variable A.

A causal model $\langle M, V \rangle$ has two components: a set M of structural equations and a set V of literals. The literals tell us which events occur and which do not. $A \in V$ means that the event whose occurrence is claimed by A does in fact occur. $\neg A \in V$, by contrast, means that this event does not occur. For ease of notation, we use upper case Latin letters for both events and propositions that a corresponding event occurs. Proposition A simply means that event A occurs.

A structural equation tells us whether an event occurs given the occurrences and non-occurrences of certain other events. Suppose A is a propositional variable and ϕ a Boolean propositional formula in which A does not occur. Further, ϕ is neither a contradiction nor a logical truth, and may not have vacuous occurrences of a variable. Then

$$A = \phi$$

is a structural equation. A variable B occurs vacuously in a formula ϕ iff there is a formula ϕ' which is logically equivalent to ϕ such that B does not occur in ϕ'. For example, B occurs vacuously in $D \land (B \lor \neg B)$.

The equation $A = \phi$ tells us whether the event represented by A occurs depending on the combination of events and absences described by ϕ. We may regard ϕ as a truth function. Its arguments represent events and absences. Its truth value determines whether A or $\neg A$. The meaning of this determination is intended to be causal: the combination of events and absences described by ϕ causally determines whether or not the event A occurs.

Causal models are thus built on the basis of a propositional language. Such a language comes with a definite set of propositional variables. The set M of structural equations of a causal model $\langle M, V \rangle$ must satisfy one important constraint. For each propositional variable A of a causal model, there is at most one structural equation $A = \phi$. We call $A = \phi$ the structural equation of A.

For now, we take structural equations to represent elementary causal dependences. The causal meaning of a structural equation is taken as primitive without further analysis. We did not say anything as to when such an equation is true relative to the world. Nor did we explain when we have reasons to believe a structural equation. These questions will be addressed at a later stage. In the synthesis of Part I and II, we will explain eventually how our broadly reductive analysis of causation may serve as conceptual foundation of causal models. This includes an account of when we are justified to accept a structural equation in a causal model.

Let us have a look at a causal model of a concrete causal scenario. Figure 1 represents a causal scenario of preemption in terms of a neuron diagram.

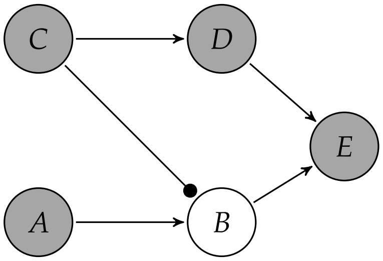


Figure 1: Early preemption

If neuron C is active, it triggers the activation of D, which in turn triggers E. C is considered a genuine cause. Neuron A, by contrast, is considered only a backup cause, at least as long as C is active. There is a causal path from A to E as well, but this path is not active, as it were, since it is interrupted by C. C inhibits activation of B. If C was not active, A would activate B, which in turn would cause E to become active. In the actual scenario, however, B is preempted. These causal dependences may be represented by the following structural equations:

$$\begin{aligned} D &= C \\ B &= A \wedge \neg C \\ E &= D \vee B. \end{aligned}$$

For readability, we represent causal models also in two-layered boxes. The first box contains the structural equations, while the literals are in the second box. The causal model of the preemption scenario, for example, is given by the following boxes:

$$\begin{array}{|l|} \hline D = C \\ B = A \wedge \neg C \\ E = D \vee B \\ \hline C, A, D, \neg B, E \\ \hline \end{array}$$

It remains to discuss two explanations of structural equations which go beyond taking such equations as elementary and primitive causal dependences. Hitchcock (2001, p. 274) understands a structural equation in terms of counterfactuals. Such an interpretation leaves the question of reduction open. If there is an analysis of counterfactuals in terms of non-causal notions, there might be a reductive analysis of causation in terms of causal models. But it is, to say the least, controversial whether a non-causal analysis of counterfactuals is feasible. Another problem is to distinguish true counterfactuals with a causal meaning from those without such a meaning. A case in point are counterfactuals which hold in virtue of some relation of grounding. For example, there are relations of counterfactual dependence between the diameter of a spherical object and the volume of that object. If the diameter had been greater, the volume would have been greater as well, and vice versa. While this type of problem is well known, it proved very difficult to exactly characterize counterfactuals which have a causal meaning—without using causal notions or some notion of metaphysical grounding at the same time. We will discuss more specific problems of Lewis's (1973a) reductive, counterfactual analysis in Chapter 10.

Halpern and Pearl (2005, p. 847) explain structural equations by reference to mechanisms and laws. They accept that such an explanation does not give us an analysis of structural equations in terms of non-causal notions. Even if we had some viable account of laws of nature in place, there would remain two problems. First, some laws of nature do not have a straightforward causal interpretation. The ideal gas law is a case in point. Second, even if a law is commonly interpreted in a causal way, it is not always possible to read the direction of causation off the notation of the law. A case in point is Newton's second equation $\mathbf{F} = m \cdot \mathbf{a}$. We typically think that forces cause accelerations, but we cannot read off this interpretation from the equation alone. A non-causal explanation of structural equations is hard to come by and for that reason we assume they express elementary causal dependences.

1.2 Semantics

The second component of a causal model $\langle M, V \rangle$ has a semantic role, even though it is given by a set of sentences. A set V of literals encodes a truth-value assignment to the propositional variables which occur in the equations in M. We can think of V as valuation of these variables. If $A \in V$, then A is assigned the truth value true. If $\neg A \in V$, then A is assigned the truth value false. If neither $A \in V$ nor $\neg A \in V$, the truth value of A is not directly specified and may be indeterminate.

We can now make the formal semantics of causal models explicit. Suppose V is complete in the sense that it contains a literal for each propositional variable which occurs in the structural equations. When does a such a complete set V of literals satisfy a propositional formula ϕ? We define a satisfaction relation using the semantics of classical propositional logic. Where $\models_{CL}$ stands for the satisfaction relation of propositional logic, we say:

$$V \models \phi \text{ iff } V \models_{Cl} \phi. \tag{$V \models \phi$}$$

In words, V satisfies ϕ iff the set V entails ϕ in the sense of classical propositional logic. Likewise, we define satisfaction for structural equations as follows:

$$V \models A = \phi \text{ iff, } V \models_{Cl} A \text{ iff } V \models_{Cl} \phi. \tag{$V \models A = \phi$}$$

In simpler terms, V satisfies the structural equation $A = \phi$ iff both sides of the equation have the same truth value on the valuation specified by V. Furthermore, we say that V satisfies a set M of structural equations just in case V satisfies each element of M:

$$V \models M \text{ iff } V \models A = \phi \text{ for all } A = \phi \text{ in } M. \tag{$V \models M$}$$

For sets Γ which contain structural equations and propositional formulas, entailment is understood in the standard way: $\Gamma \models \phi$ iff ϕ is satisfied by any complete valuation V which satisfies all members of Γ. These concepts at hand, we can define the entailment relation for causal models $\langle M, V \rangle$:

$$\langle M, V \rangle \models \phi \text{ iff } M \cup V \models \phi. \tag{$\langle M, V \rangle \models \phi$}$$

$\langle M, V \rangle$ thus entails ϕ iff ϕ is satisfied by any complete valuation which satisfies both M and V. Notice that the set V in a causal model $\langle M, V \rangle$ may well be incomplete—in the sense that it does not contain a literal for each propositional variable which occurs in the structural equations.

Now that we know the semantics of causal models, we can explain how suspension of judgement works for such models. Suppose $\langle M, V \rangle$ is a causal model and L a literal in V. So L is true on this model. To suspend judgement about L, we need to find a submodel $\langle M', V' \rangle$ of $\langle M, V \rangle$ such that L is neither true nor false on this submodel. Put differently, a causal model $\langle M', V' \rangle$ is uninformative on L iff $\langle M', V' \rangle$ entails none of L and $\neg L$. We suspend judgement on L by finding such a submodel. If L is in V, $\langle M, V \rangle$ cannot be uninformative about L.

Our epochetic analysis begins with looking for a causal model which is uninformative on the candidate cause and its effect. For brevity, we will also speak of an *agnostic model* when referring to such a model. Suspending judgement on a structural equation is needed only for certain causal scenarios. By default, we begin with suspending judgement on literals which are in V. Only if our analysis fails to recognize a presumed cause, we consider causal models $\langle M', V' \rangle$ such that $M' \subset M$.

In the scenario of preemption, for example, we can obtain two agnostic models without suspending judgement on any structural equation: $\langle M, \emptyset \rangle$ and $\langle M, \{\neg B\} \rangle$ are uninformative on the presumed cause C and its effect E. Neither model entails C, E or a negation of C or E. Further agnostic models come into play when we suspend judgement on structural equations: for any $M' \subset M$, $\langle M', \emptyset \rangle$ and $\langle M', \{\neg B\} \rangle$ are uninformative on C and E. However, we will constrain suspension of structural equations in ways to be explained below.

Note that our semantics of causal model is completely analogous to the semantics of classical propositional logic. The semantics of $=$, in particular, does not differ from the semantics of the classical biconditional $\leftrightarrow$. A structural equation $A = \phi$ is satisfied by a valuation V (given by a set of literals) iff both sides of the equation have the same truth value on the valuation V. This means that our structural equations are symmetric just like classical biconditionals: we can infer the left-hand side from the right-hand side and the other way around.

The two directions of an inference with a structural equation correspond to two different types of causal reasoning. When we infer A from $A = \phi$ and ϕ, our inference goes from causes to an effect. We call such inferences *forward-directed*. By contrast, when we infer ϕ from $A = \phi$ and A, our inference goes from an effect to a sentence about its causes. We call such inferences *backward-directed*. We could also call them *abductive*.

Forward-directed inferences are aligned with the direction of causation since they go from causes to an effect. This does not imply that such inferences are always aligned with the direction of time. We use the concepts of forward and backward in a causal, not in the temporal sense here. Time does not come into play in Part I, but will be important for Part II.

In an analysis of causation, one is particularly interested in the forward-directed consequences of a causal model $\langle M, V \rangle$. In line with Halpern

(2000) and Pearl (2009), we introduce an operation of intervention in order to separate the forward-directed consequences from the backward-directed ones. Our operation of intervention is basically the syntactic counterpart of interventions as defined there.

1.3 Interventions

We have seen that the semantics of structural equations is symmetric, while the intended meaning of such an equation favours a specific direction of determination. If $A = \phi$ is a structural equation, then ϕ determines A but not the other way around. How can we account for the direction of this determination?

Suppose we want to determine the forward-directed causal consequences of the occurrence of A for a causal model $\langle M, \emptyset \rangle$. We can accomplish this by, first, removing the structural equation of A from M if there is one. Second, we add A to the set of literals V, which is given by the empty set in this example. Once the structural equation of A is eliminated from M, there is no way to infer a sentence about the causes of A anymore. Backward-directed inferences are blocked. Inferences about the effects of A, by contrast, may still be drawn if M contains an equation which has an occurrence of A on the right-hand side.

The case where the valuation V in a causal model $\langle M, V \rangle$ is not empty is more complex. We need to intervene, first, by the valuation V and then by the premise A in order to determine the forward-directed consequences of A. Hence, we need to define interventions by sets of literals. Suppose I is such a set. Let us denote interventions by an operator $[\cdot]$ which takes a causal model $\langle M, V \rangle$ and a set I of literals, and returns a causal model—the submodel of $\langle M, V \rangle$ after the intervention by I. The intervention by a set of literals is defined as follows:

$$\langle M, V \rangle \, [I] = \langle M_I, V \cup I \rangle \qquad\qquad (\langle M, V \rangle \, [I])$$

where

$$M_I = \{(A = \phi) \in M \mid A \notin I \text{ and } \neg A \notin I\}.$$

M_I is the subset of M which contains each structural equation $A = \phi$ whose variable A is not evaluated by any member of I. After intervening by I on the causal model $\langle M, V \rangle$, the set I becomes part of the valuation of the resulting submodel. Note that the resulting submodel $\langle M_I, V \cup I\rangle$ is again a causal model, consisting of a set of structural equations and a set of literals. Iterated interventions are thus well defined.

Interventions may well result in inconsistent causal models. Even if the original causal model $\langle M, V\rangle$ is consistent, $M_I \cup V \cup I$ may well be inconsistent. There is, however, no reason for concern. In our epochetic analysis, we need interventions only as part of the definition of the conditional $\gg$. And this definition sidesteps such inconsistencies by means of a prior operation of suspending judgement.[1]

Let us explain causal reasoning by means of interventions on a causal model $\langle M, V\rangle$ which is agnostic as regards C and E. We are interested in the forward-directed consequences of C in this model. These are obtained by the classical consequences of $\langle M, V\rangle[V][C]$. To be precise, a sentence ϕ is a forward-directed consequence of C in $\langle M, V\rangle$ iff ϕ is entailed by $\langle M, V\rangle[V][C]$. The causal model $\langle M, V\rangle[C]$, by contrast, may still contain consequences of C which are obtained by a combination of backward and forward-directed causal reasoning. This is why iterated interventions are needed to determine the forward-directed consequences of C.

Note that $\langle M, V\rangle[V]$ differs from $\langle M, V\rangle$ in that it does not contain the structural equations of those variables which are evaluated by V. If, for example, $\neg A$ is in V and M contains a structural equation $A = \phi$, then this equation is not a structural equation of $\langle M, V\rangle[V]$. In more formal terms, $\langle M, V\rangle[V]$ equals $\langle M_V, V\rangle$, and M_V does not contain the equation $A = \phi$. Obviously, the valuation V remains unchanged when intervening by V.

Recall why the distinction between $\langle M, V\rangle[V]$ and $\langle M, V\rangle$ is important. $\langle M, V\rangle[V]$ gives us the forward-directed consequences of the set V in the context of M. All sentences ϕ which are entailed by $\langle M, V\rangle[V]$ are forward-directed consequences of $V \cup M$. The causal model $\langle M, V\rangle$, by contrast, entails all consequences of $M \cup V$ without any discrimination between forward and backward-directed causal reasoning. Entailment has been defined along the lines of classical logic in the above section.

[1] Counterfactual interventions may nonetheless be defined using the present framework. See Andreas and Günther (2025b), which is based on Appendix A.

1.4 Natural Deduction

We have now seen how interventions enable us to consider the direction of determination in structural equations. This is the key result: once we have intervened with the set V of literals on a causal model $\langle M, V \rangle$, the semantics of classical logic gives us the forward-directed consequences of this model. *Forward-directed* means that the direction of reasoning is aligned with the direction of causation: it goes from causes to effects.

These considerations apply to deductive reasoning with structural equations as well: once we have intervened with the set V of literals on a causal model $\langle M, V \rangle$, we can use the classical inference rules in order to draw forward-directed conclusions from this model. Such conclusions are sentences about events and absences which are effects of causes asserted by V.

Suppose $\langle M, V \rangle$ is a causal model which is uninformative on the candidate cause C and its presumed effect E. We are interested which conclusions we can infer—by means of deductive reasoning—from $\langle M, V \rangle$ after an intervention by V and C. Now, it is surprisingly simple to define a deductive system for forward-directed inferences from a causal model $\langle M, V \rangle [V][C]$. The following rules of natural deduction suffice for deductive reasoning with structural equations:

(1) The inference rules for the Boolean connectives $\wedge$, $\vee$, and $\neg$.

(2) Two inference rules for the equality symbol $=$.

The latter rules are as follows:

$$\frac{A = \phi \qquad \phi}{A} \qquad\qquad \frac{A = \phi \qquad \neg\phi}{\neg A}.$$

These two inference rules are analogous to Modus Ponens in classical logic. From $A = \phi$ and ϕ, we can infer A. From $A = \phi$ and $\neg\phi$, we can infer $\neg A$. For causal models of the form $\langle M, V \rangle [V]$, these inference rules may be used without further restrictions. The intervention by V ensures that we derive only causally forward-directed inferences from the set V of premises. Likewise for causal models $\langle M, V \rangle [V][C]$. No further inference rules are needed to capture the forward-directed conclusions from a causal model. If there is a deduction of ϕ from $\langle M, V \rangle$, we write $\langle M, V \rangle \vdash \phi$.

One word on the distinction between formal and semiformal deductions is in order. Mathematical and scientific reasoning is virtually always semiformal in that formal equations and symbols are used together with nonformal elements and intuitions. Some intermediate conclusions are left out. Certain principles of inference and some premises remain implicit. Only deductions in a formal logical system are purely formal and syntactic. Even the metatheory of a logical system is almost always studied in a semiformal fashion.

For causal models we have both options. First, we can use the system of inference rules outlined here in order to draw inferences from a causal model $\langle M, V\rangle[V]$. This way, we obtain fully formal deductions. Second, we can draw inferences in a semiformal fashion, which implies that some inferential steps may be left out. Semiformal deductions are rarely purely syntactic. When we say, for example, that A, B and the equation $E = A \wedge B$ determine E to be true, a semantic understanding of this inference is in the background. Truth is not a syntactic concept.

For our analysis to work, we need to make only those inferential steps explicit which conclude with a literal. The deduction of the effect from the genuine cause is easy to recognize—without considering the details of a fully formalized deduction—in all causal scenarios studied in this book. But it will be important to distinguish between relations of semantic entailment and deductions from a causal model. A deduction makes at least some inferential steps explicit, while an entailment relation does not. Certain subtle properties of genuine causes will become transparent only when we study the inferential pathways from the candidate cause to the effect. It does seem to matter how we can infer the effect from a candidate cause. Mere entailment is insufficient, as will be argued in greater detail in Section 4 and 5.

We develop further details of this logic of causal models in Appendix A. In this section, we have confined ourselves to those concepts which are essential for our analysis of causation to follow. Key results obtained in the appendix remain important and valid for the simplified account, though. This applies, in particular, to theorems about soundness and completeness for the deductive system just expounded. The following theorem holds.

Theorem 1. Let ϕ be a Boolean propositional formula. There is a deduction of ϕ from $\langle M, V\rangle$ iff ϕ is entailed by $\langle M, V\rangle$. In symbols, $\langle M, V\rangle \vdash \phi$ iff

$\langle M, V \rangle \models \phi$.

Notice that $\langle M, V \rangle$ may be obtained from a causal model by an intervention with V such that the direction of causation is considered. The proof is obvious from corresponding theorems in Appendix A.

2 Causal Graphs

Any causal model $\langle M, V \rangle$ gives rise to a single causal graph. Such a graph tells us which variables are connected to which other variables by a relation of direct causal determination.

A causal graph has the general form of a directed graph. Such a graph is commonly represented by an ordered pair $\langle \mathcal{N}, \mathcal{E} \rangle$ such that $\mathcal{N}$ is a set of nodes and $\mathcal{E}$ a set of directed edges. For a causal graph, the set of nodes contains all the variables of the respective causal model. The edges stand for relations of direct causal determination. Let us study a concrete example. Recall the preemption scenario from the previous section with its distinction between a genuine cause C and a backup cause A. Figure 1 represents the neuron diagram of this scenario.

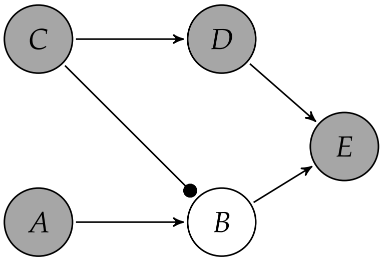


Figure 1: Early preemption

If we have a representation of a causal scenario in terms of a neuron diagram, the corresponding causal graph can be read off the diagram. In the preemption scenario, the nodes are obviously given by the set $\{A, B, C, D, E\}$. The edges of the graph are as follows: (A, B), (B, E), (C, D), (C, B), and (D, E). More intuitively, we can also write $\mathcal{E} = \{A \to B, B \to$

$E, C \rightarrow D, C \rightarrow B, D \rightarrow E\}$. Here is a diagram of the causal graph for the preemption scenario:

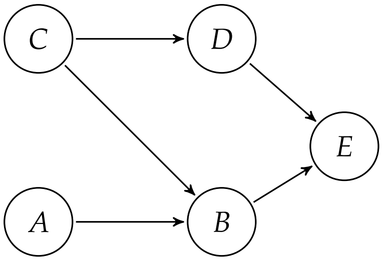


Figure 2: Causal graph of early preemption

By now, it should be obvious how we can read the causal graph off the neuron diagram of a causal scenario. Note that the edges stand for relations of direct causal determination. $C \rightarrow D$, for example, is an edge since C directly causally determines D. Both activation and inhibition are considered relations of direct determination. An edge between A and B, however, does not imply that A is an *actual* direct cause of B. We need to distinguish between direct causal determination and direct actual causation, as will become more obvious shortly.

In addition to neuron diagrams, we can read the causal graph off the set M of structural equations of a given causal model. In general, $A \rightarrow B$ is an edge of the causal graph iff there is an equation σ in M such that σ has the logical form $B = \phi$ and A occurs in ϕ. To see how this rule is working, recall the structural equations of the preemption scenario from the previous section:

$$D = C$$
$$B = A \wedge \neg C$$
$$E = D \vee B.$$

Applying the just explained rule yields the same set of edges as the neuron diagram does. $A \rightarrow B$, for example, is an edge since there is an equation $B = \phi$ such that A occurs on the right-hand side of this equation.

Note that the causal graph of a causal model does not depend on the actual values of the variables. Such values completely drop out of the consideration when it comes to causal graphs. Consequently, the edges of the graph do not stand for relations of actual causal determination. In our example, B is not active, and so does not activate E. It is not a direct cause of E in the actual causal scenario. We still say that B stands in a relation of direct causal determination to E since the value of B determines—together with the value of D—the value of E.

Finally, we need to study a few concepts concerning pathways in a causal graph. A directed path is a sequence of nodes which are connected by edges which are aligned. A directed path is thought to begin with the first element of the sequence and to end with the last. Multiple occurrences of one and the same node are not allowed. For example, $\langle A, B, E\rangle$ is a directed path in the causal graph of our example. We can also write $A \rightarrow B \rightarrow E$ for this path.

An undirected path, by contrast, is a sequence of nodes which are connected by edges which may not be aligned. As with directed paths, no node must occur more than once in an undirected path. $A \rightarrow B \leftarrow C$, for example, is an undirected path in the graph of the preemption scenario. The sequence $C \rightarrow B \rightarrow E \leftarrow D \leftarrow C$, by contrast, is neither a directed nor an undirected path in this graph.

Suppose $\langle A, \ldots, A\rangle$ is sequence of nodes which are connected by edges which are aligned. Such a sequence is called a *cycle*. It starts from a node and goes back to it along a sequence of edges which are aligned. A causal graph is called *acyclic* iff it does not contain any cycles. All scenarios studied in this book have causal models whose graph is acyclic. While it may be controversial whether a given event A can cause itself, we assume that no causal graph contains an edge of the type $A \rightarrow A$.

It is sometimes helpful to think of causal relations in terms of ancestry. Suppose there is a directed path from A to B in the causal graph of M. Then A is called an *ancestor* of B. And B is called a *descendant* of A. The idea is that ancestors are potential causes of their descendants.

3 Inferential Pathways

The inferential analysis to follow exploits an important graph-theoretic property of deductions in causal models: deductive reasoning with structural equations always proceeds in a stepwise fashion along the edges of the causal graph of M: we cannot 'jump' from one literal to another if there is no directed edge between the variables of the two literals. Intermediate conclusions are always needed in this case.

Proposition 1. Suppose $\langle M, V\rangle$ is uninformative on the literals L_A and L_B. Further, suppose $\langle M, V \cup \{L_A\}\rangle \vdash L_B$. Then, for any deduction of L_B from $\langle M, V \cup \{L_A\}\rangle$, there is an undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ ($n \geq 0$) of variables such that, if $n > 0$, the deduction contains an intermediate conclusion for each variable $D_1, \ldots, D_n$.

Henceforth, L_A stands for a literal of the variable A. Likewise for other variables. While a deduction always goes along an undirected path, it may not be aligned with a directed path in the causal graph. If, however, we intervene with the valuation V and the premise L_A, then the deduction of another literal L_B always contains an inferential path which is aligned with a directed path in the causal graph.

Proposition 2. Suppose $\langle M, V\rangle$ is uninformative on the literals L_A and L_B. Further, suppose $\langle M, V\rangle[V][L_A] \vdash L_B$. Then, for any deduction of L_B from $\langle M, V\rangle[V][L_A]$, there is a directed path $\langle A, D_1, \ldots, D_n, B\rangle$ ($n \geq 0$) of variables such that, if $n > 0$, the deduction contains an intermediate conclusion for each variable $D_1, \ldots, D_n$. Such a directed path exists in the causal graph of $M_{V \cup \{L_A\}}$.

This proposition implies that any deduction of a literal L_B from $\langle M, V\rangle[V][L_A]$ is causally forward-directed: the variable B is then a descendant of the variable A. This does not hold for deductions from $\langle M, V \cup \{L_A\}\rangle$. To understand the reason for this difference, recall that an intervention by V and L_A does not only affect the valuation of the causal model, but also the set of structural equations (see Section 1). The causal model $\langle M, V\rangle[V][L_A]$ contains only structural equations of those variables which do not have occurrences in the literals in $V \cup \{L_A\}$.

4 The Inferential Analysis

Our inferential analysis aims to reconstruct how a candidate cause brought about a given effect. The key idea is that each inferential step to a literal may be interpreted as a section in a causal process which was started by the candidate cause. We express this condition by a constraint on the inferential pathways leading from the candidate cause to the effect: any inferential step to a literal, made by a structural equation, must depend on the assumption of the candidate cause. Inferential paths which satisfy this condition are called *active*.

Definition 1. Active Path
Let $\langle M, V \rangle$ be a causal model, which is uninformative on the literals C and E. There is an active path leading from C to E in $\langle M, V \rangle$ iff E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal—by a structural equation—depends on C.

This notion of active path is defined in terms of relations of inferential dependence among literals. Such relations we claim tell us whether or not a candidate cause is genuine. What does it mean that an inferred literal depends on another?

A literal E inferentially depends on a literal C in a given deduction iff C is used, directly or indirectly, to infer E. In this sense, C may be needed to infer E in a specific deduction. Suppose we have inferred E from the set $\langle M, V \rangle[V][C]$ in a stepwise fashion such that the inferential steps which rely on a structural equation are made explicit. This gives us a deduction of E from $\langle M, V \rangle[V][C]$, which may be carried out in a formal system or in a semiformal fashion. In any case, it is instructive to consider the two inference rules for structural equations:

$$\frac{A = \phi \qquad \phi}{A} \qquad\qquad \frac{A = \phi \qquad \neg\phi}{\neg A}.$$

For causal models $\langle M, V \rangle[V]$, these rules of inference may be used without further restrictions. The intervention by V ensures that we derive only causally forward-directed inferences from the set V of premises.

To infer a literal A using the equation $A = \phi$, ϕ must be a literal in $V \cup \{C\}$ or it must be inferred from $V \cup \{C\}$ using other equations in M. Likewise, to infer a literal $\neg A$ using the equation $A = \phi$, $\neg\phi$ must be a literal in $V \cup \{C\}$ or it must be inferred from this set. The two cases are completely analogous, which is why we only need to consider one.

Since, by assumption, C is needed for the deduction of E, there is at least one inferential step in this deduction in which C is used as a premise directly. We may, for example, infer a Boolean formula from C, such as $C \vee A$

or $C \wedge A$. The latter inference depends, of course, also on A being given. Or we may infer another literal B from C, provided M contains an equation $B = C$. All these inferences depend on the literal C. The key condition of our definition of active path is satisfied for all these inferences.

Suppose now that the equation $A = \phi$ is used to infer A, while ϕ is not in the set $V \cup \{C\}$. Then ϕ must be inferred from $V \cup \{C\}$, possibly using the equations in M. Put differently, ϕ is an intermediate conclusion on the way to the final conclusion E. It is intermediate between $V \cup \{C\}$ and E. We arrive at the intermediate conclusion ϕ by way of one or more inferential pathways, all of which start from $V \cup \{C\}$. Note that these pathways go along the directed edges which connect the literals in $V \cup \{C\}$ with E in the causal graph of M. The crucial point is that at least one of the pathways must have an occurrence of C as a formula. If this condition is satisfied for all inferential steps to a literal, then we say there exists an *active path* from C to E. Otherwise, no such path exists.

It may be interesting to see how we can define the relation of dependence between two literals in terms of natural deduction. Suppose we use a tree representation of natural deduction. Then every conclusion—be it intermediate or final—has one or more branches. Each branch corresponds to a sequence of formulas. At the top of each branch, we have a premise or an assumption of a subproof. Suppose ϕ is inferred in a natural deduction proof of E from $\langle M, V \rangle [V][C]$ in order to infer A using the structural equation $A = \phi$. Then we say that the inference of A depends on all literals which occur as formula in at least one branch of ϕ. The system of the branches of ϕ makes up the derivation of ϕ.[2]

One more clarification is needed concerning the relation of dependence among literals. We have explained what it means that a literal depends on another with respect to a given deduction of the effect from the candidate cause. Such a deduction must satisfy an important constraint, which has been tacitly assumed so far: it must not contain redundant inferential steps. The inferential pathways must not contain sections which may be eliminated without thereby affecting the correctness of other inferential steps. In other words, the deduction of the final conclusion E must not contain sections which may be eliminated in such a manner that the remaining

[2] For an exposition of classical logic using the tree representation of natural deduction, see, e.g., Zach (2021).

derivation is a correct deduction of E. A non-redundant deduction thus must satisfy a certain minimality constraint.

To give a simple example of a redundant deduction, suppose we infer E from A using the equation $E = A$. Obviously, E depends on A then. Now, suppose A and B are literals in $V \cup \{C\}$. From A and B we could then infer the conjunction $A \wedge B$, from which we could infer A again. Then we infer E using the equation $E = A$. There then is an inferential pathway leading to the inferential step to E on which B occurs. From this, however, we should not infer that E depends on B since the inferential steps to $A \wedge B$ and back to A are redundant. They can be eliminated without affecting the correctness of the inferential step to E.

When we say that a literal E depends on another literal A, then this means that A is needed to infer E in a given deduction. In practice, it is not difficult at all to work out a non-redundant deduction. For virtually all scenarios to be discussed in the following chapters, there is an agnostic model with an obvious deduction of the effect from the candidate cause such that this deduction is not redundant. This holds for both presumably genuine and non-genuine causes.

It is finally worth noting that there are two principal ways how a literal may be inferred by a structural equation. First, directly from a structural equation $A = \phi$, and ϕ or $\neg\phi$. Second, indirectly by a subproof in which a structural equation is used to infer a literal directly. Since causally meaningful deductions with subproofs rely on direct deductions with structural equations, inferential dependence concerning subproofs may be understood on the basis of inferential dependence in direct deductions. This will become obvious in the discussion of causal scenarios to follow.

Such is the relation of inferential dependence between two literals, which underlies our notion of active path. We are now in a position to set forth our core analysis. In the first part, we define an epochetic conditional:

Definition 2. $\langle M, V \rangle \models C \gg E$
Let $\langle M, V \rangle$ be a causal model. $\langle M, V \rangle \models C \gg E$ iff there are $V' \subseteq V$ and $M' \subseteq M$ such that

(1) $\langle M', V' \rangle$ is uninformative on C and E.

(2) There is an active path from C to E in $\langle M', V' \rangle$.

(3) All the structural equations of C's descendants are in M'.

The second part says that C is a genuine cause of E iff C and E are actual, and there is an active path from C to E in an agnostic causal model which contains the structural equations of all descendants of C. In more formal terms:

Definition 3. Cause
Let $\langle M, V \rangle$ be a causal model such that $V \models M$. C is a cause of E relative to $\langle M, V \rangle$ iff

(C1) $\langle M, V \rangle \models C \wedge E$, and

(C2) $\langle M, V \rangle \models C \gg E$.

In essence, our analysis says that there is an active path from a genuine cause to its effect in an agnostic model. Such an agnostic model must contain the structural equations of the descendants of the candidate cause. Pseudo causes don't have an active path to the effect in an agnostic model with this property. Some pseudo causes don't have an active path at all, some have an active path in an agnostic model which lacks a structural equation of the candidate cause.

The constraint on agnostic models is motivated by results about inferential pathways from the previous section: for any deduction of E from $\langle M', V' \rangle [V'][C]$, there is a directed path from C to E in the respective causal graph. By requiring that M' contains the structural equations of all descendants of C, we ensure that all forward-directed inferential connections between the candidate cause and its effect are preserved when we suspend judgement on C and E.

The condition that $\langle M', V' \rangle$ is uninformative on C and E may be understood in terms of deductions or entailment. We can say, for example, that $\langle M', V' \rangle$ is uninformative on C iff there is no deduction of C or $\neg C$ from $\langle M', V' \rangle$. Likewise for E. Thanks to soundness and completeness of our deductive system for structural equations, it does not matter which way we understand the notion of an agnostic model. The notion of active path, by contrast, does not seem to have an obvious semantic counterpart in terms of entailment. Inferences proceed in a stepwise fashion, semantic valuations do not.

The present analysis captures a large range of causal scenarios. Only one refinement will be needed before we can state our final analysis of the notion of cause. It is the infamous problem of isomorphic causal models which will make a refinement of the analysis necessary later on.

It is striking that, so far, our analysis merely explains what it is for a causal model $\langle M, V \rangle [V]$ to have an active path from C to E. The active path itself remains unspecified. Our analysis is, in principle, applicable without any further explanation of what an active path is. However, it will be easier to grasp if furnished with a more direct explanation of the notion of active path. Such an explanation will also ease the discrimination between genuine and non-genuine causes. We will work it out in the next section.

5 Inferential Networks

To get a better grip on active paths, we introduce the concept of inferential network between two literals in a deduction. This concept is to make dependence relations among literals explicit whenever we can infer E from $\langle M, V \rangle [V][C]$. An inferential network is a graph $\langle \mathcal{L}, \mathcal{D} \rangle$, the nodes of which are literals. More specifically, $\mathcal{L}$ contains each occurrence of a literal in a given deduction whenever this literal is inferred. If one and the same literal is inferred more than once, we need to distinguish all of these occurrences. In addition, $\mathcal{L}$ is defined by the condition that C is a member of it. The inferential network $\langle \mathcal{L}, \mathcal{D} \rangle$ rarely coincides with the causal graph of M.

The edges in $\mathcal{D}$ are defined as follows. A directed edge $A \rightarrow B$ is in $\mathcal{D}$ iff E directly depends on A. To give a simple example, if E is inferred from A using the equation $E = A$, then E directly depends on A. Moreover, if E is inferred from A and B using the structural equation $E = A \wedge B$, then E directly depends on A and B. If, however, we infer E from $E = A \vee B$, E might not depend on A at all. For example, if $A \vee B$ has been inferred from B, then the inferential step to E may well be independent of A. This is why we say that a literal E depends on another literal if the latter occurs as a formula—rather than as a mere subformula—on the inferential pathways to the conclusion E.

In more general terms, E directly depends on a literal A iff there is no intermediate conclusion of a literal L on the inferential pathway from A to

E. Again, a bit of natural deduction is needed to make this idea fully precise. Suppose E is inferred by a structural equation $E = \phi$. And the literal A occurs—as a formula—in a branch which makes up the derivation of E. We say that E directly depends on A iff the section of the branch from A to E has no occurrences of any other literal as a formula. This condition amounts to there being no intermediate conclusion in the form of a literal between an occurrence of A and an occurrence of B as formulas in the branch. Recall that one and the same literal may occur more than once in a deduction.

The careful reader may have noticed that we leave out premises in V and assumptions of subproofs from the representation of dependences among literals. Why so? The simple answer to this question is that these dependences are not needed in order to discriminate between genuine and non-genuine causes. The discrimination becomes even easier when we leave out said premises. And the active path of a genuine cause can be read off the inferential network directly. This will be shown shortly.

Let us now study the inferential networks of some concrete scenarios. The networks of genuine causes are almost always quite simple. Take the preemption scenario and the agnostic model $\langle M, \emptyset \rangle$: $C \rightarrow D \rightarrow E$ is an inferential network from C to E for the genuine cause in this model. This is easy to verify with the neuron diagram of the preemption scenario considered:

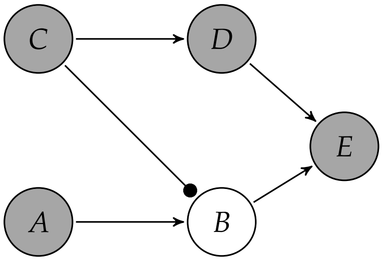


Figure 1: Early Preemption

The inferential network for the preempted cause is more interesting. To infer the effect from the preempted cause in the causal model $\langle M, \emptyset \rangle$, reasoning by cases is needed. We need to distinguish two cases: first, C is true, and, second, C is not true. If C is true, we can infer E via D. If C is not true, we can infer B from A and $\neg C$, which gives us E. Thus we have shown that E must be true no matter which assumption we make about C, provided A is true. The inferential network of this deduction may be represented as follows:

$$A \to B \to E' \to E \leftarrow E'' \leftarrow D.$$

The literal E has three different occurrences in this network since it is inferred three times. First, E is inferred in the two subproofs, respectively. Then as the final conclusion of the deduction. Clearly, this inferential network is not an active path since it violates the crucial condition that all inferred literals depend on the candidate cause A. Both D and E'' violate this condition.

A remarkable property of our analysis is that we can now distinguish between preempted and genuine causes even in the absence of intermediate events—intermediate between the candidate cause and the effect. Take the following structural equation:

$$E = C \vee (A \wedge \neg C).$$

Suppose C and E are actual. The structural equation alone may or may not convince us that C is a genuine cause, while A is not. We will get to know a concrete scenario which we think should convince us of these judgements in the next chapter. Clearly, the only agnostic model which preserves the structural equations of C's descendants is $\langle M, \emptyset \rangle$. Obviously, C has a straightforward active path: $C \to E$.

The preempted cause A, by contrast, does not have such a path. Again, the deduction of the effect for the preempted cause is more complex since reasoning by cases is needed. First, we assume C from which we can infer E. If we assume $\neg C$, we can infer from this and A that E. Hence, whatever assumption we make about C, E must be true. The inferential network of this deduction is as follows:

$$A \to E' \to E \leftarrow E''.$$

We have different occurrences of E since this literal is inferred several times in the deduction. Clearly, this network does not contain an active path since E'' is not on a directed path from A to the final conclusion E. For all we know, our analysis is the first to discriminate between genuine and preempted causes without intermediate events. No other analysis of causation in terms of causal models has accomplished this so far.

An inferential network lets us recognize certain structural properties of a given deduction concerning the inferential dependences among literals. Such a network does not give us a complete deduction. Subproofs, for example, are not marked as such. Explicit premises other than that of the candidate cause are left out. Likewise assumptions of subproofs. An inferential network may be seen as the skeleton of a complete deduction.

Notice that both the inferential network $\langle \mathcal{L}, \mathcal{D} \rangle$ of a deduction and the causal graph of M are directed graphs in the sense of graph theory. Despite some similarities, they hardly coincide with one another. They differ for a number of reasons. First, the inferential network may have several occurrences of one and the same literal. Second, the network disregards the literals in V for simplicity (even though these literals may be used in the deduction of E). Third, the nodes of the causal graph of M are variables, while the nodes of an inferential network are literals. Only for very simple causal scenarios, the two graphs may be seen to coincide when we ignore the distinction between variables and positive literals. A case in point is the causal model with the single structural equation $E = C$.

The crucial property of inferential networks, as defined here, is that they give us complete information about all relations of dependence among all literals inferred on the way to the conclusion of the effect E as well as relations of dependence concerning the candidate cause C. In view of this, we can now give a direct explanation of the notion of active path.

Explanation 1. Active Path as Inferential Network
Let $\langle \mathcal{L}, \mathcal{D} \rangle$ be the inferential network for a deduction of E from $\langle M, V \rangle [V][C]$. This inferential network is an active path from C to E iff all nodes of this network are on a directed path from C to E in this network.

Moreover, we can give an alternative explanation of when there is an active path from a candidate cause to the effect.

Definition 4. Active Path
Let $\langle M, V \rangle$ be a causal model, which is uninformative on the literals C and

E. There is an active path leading from C to E in $\langle M, V\rangle$ iff there is a deduction of E from $\langle M, V\rangle[V][C]$ such that the inferential network of this deduction has the property that all nodes are on a directed path from C to E.

This definition is intended to be equivalent to the definition of active path in the previous section. In fact, we can prove the following proposition.

Proposition 3. Let $\langle M, V\rangle$ be a causal model which is uninformative on the literals C and E. There is a deduction of E from $\langle M, V\rangle[V][C]$ such that the inferential network of this deduction has the property that all nodes are on a directed path from C to E iff E can be inferred from $\langle M, V\rangle[V][C]$ such that any inferential step to a literal depends on C.

To understand why the equivalence holds, it is helpful to realize that (i) inferential networks give us complete information about direct and non-direct dependences among the inferred literals, including dependences with respect to C. Suppose a network contains the directed path $B \rightarrow D \rightarrow E$. This tells us that E directly depends on D, which in turn directly depends on B. Moreover, the path indicates that E depends on B, albeit not directly. It holds in general that (ii) if there is a directed path from literal C to literal D—in a network of a deduction from $\langle M, \emptyset\rangle[V][C]$—then we know that D inferentially depends on C. Finally, (iii) if literal D inferentially depends on literal C, then there is a directed path from C to D in the network of the corresponding deduction. With these observations at hand, the proof is relatively straightforward (see Appendix C for details).

Clearly, if the inferential network has the form of a sequence, then this sequence is always an active path. To be more precise about this connection:

Proposition 4. Let $\langle \mathcal{L}, \mathcal{D}\rangle$ be the inferential network of a deduction of E from $\langle M, V\rangle[V][C]$. If this network has the form of a sequence $C \rightarrow \ldots \rightarrow E$, then this sequence is an active path.

The proof of this proposition is obvious from Explanation 1. We can even state a stronger proposition: whenever the inferential network of a deduction of E from $\langle M, V\rangle[V][C]$ has the form of a sequence, then this sequence is an active path.

These observations greatly ease the application of our analysis to concrete scenarios. First, we look for a causal model which is uninformative on

C and E. Then we see if we can infer E from C such that the inferential network of this inference is a sequence from C to E. This, in a nutshell, is how our epochetic analysis will be used when it comes to recognizing genuine causes in Part I.

Speaking of an active path suggests, if not implies, that such a path has always the form of a sequence. A directed path in a graph is always a sequence of nodes. Does an active path, as defined here, always have this specific form? This is in fact the case for all genuine causes in all scenarios which are widely discussed in the literature. In the following chapters, we will not encounter a single scenario where the active path of a genuine cause has a structure other than a sequence of literals. However, there are counterexamples to the conjecture that an active path always has the form of a sequence.

Suppose two events A and B have a common cause C. Further, A and B are conjunctive causes of a common effect E. This scenario is represented by the following structural equations:

$$\begin{aligned} A &= C \\ B &= C \\ E &= A \wedge B. \end{aligned}$$

Obviously, $\langle M, \emptyset \rangle$ is a causal model which is uninformative on C and E. There is an obvious deduction of E from $\langle M, \emptyset \rangle [C]$. Its inferential network is as follows:

$$C \rightarrow A \rightarrow E \leftarrow B \leftarrow C.$$

Clearly, this inferential network is an active path in the sense of Explanation 1. All nodes are on a directed path from the candidate cause to the effect. This network, however, does not have the form of a sequence. The example tells us that not all active paths have the form of a sequence. At the same time, notice that the graph $C \rightarrow A \rightarrow E \leftarrow B \leftarrow C$ equals the set-theoretic union of the graph $C \rightarrow A \rightarrow E$ with the graph $C \rightarrow B \rightarrow E$. Figuratively speaking, our network $C \rightarrow A \rightarrow E \leftarrow B \leftarrow C$ is a *bundle* of paths, all of which have the form $C \rightarrow \ldots \rightarrow E$.

More generally, it is easy to show that any inferential network which is an active path has the form of a sequence $C \rightarrow \ldots \rightarrow E$ or equals the union of such sequences. Hence, we can say that some active paths have a more complex structure, which contains several subpaths from C to E. Of course, we are stretching the meaning of the word *path* a bit when we call $C \rightarrow A \rightarrow E \leftarrow B \leftarrow C$ an *active path*. Strictly speaking, this network is not a path in the sense of graph theory. The benefit of our terminology is overall simplicity and accessibility. All active paths in causal scenarios which have been widely discussed in the literature have the form of a sequence. This will be shown in the following chapters.

We have thus pinned down our analysis of causation to a relatively simple criterion concerning inferential networks: all nodes of such a network need to be on a directed path between the candidate cause and the effect. Applying our analysis therefore becomes straightforward as soon as we have the inferential networks laid out. A network tells us in which ways we can infer the effect from the candidate cause in an agnostic model.

In principle, we have to consider the networks of several deductions for a given candidate cause. Moreover, several agnostic models may have to be considered. In practice, however, there is often an obvious agnostic model and an obvious deduction of the effect from the candidate cause. They are obvious from the causal graph of the respective scenario. This will be shown with reference to a larger number of causal scenarios in the chapters to follow. Variations of what will strike us an obvious deduction either do not change the inferential network or lead to a redundant deduction. The latter type of deduction must not be considered for the recognition of genuine causes, for reasons explained in the above section.

The essence of our epochetic analysis in Part I is thus as follows. First, we look for a causal model which is uninformative on C and E. Then we see if we can infer E from C such that the inferential network of this deduction is an active path. If it is, the candidate cause is genuine. If it is not, we have to find another agnostic model for which there is a deduction with the desired property. If we do not find such a deduction, we have to try harder or show that there is none. We will say more about the latter problem in the following chapter.

Chapter 3

Classics

We have proposed an analysis of causation in the previous chapter. In the chapters to come, we test our analysis against our causal judgments and show that it agrees with the commonsensical judgments about a very wide range of causal scenarios. In this chapter, we begin with classic causal scenarios known as conjunctive causes, overdetermination, early and late preemption, trumping, prevention, and double prevention. We show that our core analysis delivers the desired judgments for these scenarios.

Causal scenarios have received much attention in the literature and for a good reason. They are an important criterion for assessing analyses of causation. An analysis is *prima facie* better if its verdicts agree with our judgments on a wider range of causal scenarios. 'Our' judgments are here the judgments of common sense. For quite a wide range of causal scenarios, the commonsensical judgments have been made explicit in the textbook-like guide to causation by Paul and Hall (2013). Throughout we take it that they provide an accurate account of our commonsensical judgments.

1 Neuron Diagrams

We follow Paul and Hall (2013, p. 10) in laying out the structure of causal scenarios by *neuron diagrams*. 'Neuron diagrams earn their keep', they write, 'by representing a complex situation clearly and forcefully, allowing the reader to take in at a glance its central causal characteristics.'[1] We introduce now simple neuron diagrams before we apply our analysis to the causal scenarios.

A neuron diagram is a graph-like representation that comes with different types of arrows and different types of nodes. Any node stands for a neuron, which fires or else does not. The firing of a neuron is visualized by a gray-shaded node, the non-firing by a white node. For the scenarios to be considered, we need two types of arrows. Each arrow with a head represents a stimulatory connection between two neurons, each arrow ending with a black dot an inhibitory connection. Furthermore, we distinguish between *normal* neurons that become excited if stimulated by another and *stubborn* neurons whose excitation requires two stimulations. Normal neurons are visualized by circles, stubborn neurons by thicker circles. The neuron diagrams to follow obey four rules. First, the temporal order of events is left to right. Second, a normal neuron will fire if it is stimulated by at least one and inhibited by none. Third, a stubborn neuron will fire if it is stimulated by at least two and inhibited by none. Fourth, a neuron will not fire if it is inhibited by at least one.

Typically, neuron diagrams are used to represent events and absences. The firing of a neuron indicates the occurrence of some event and the non-firing indicates its non-occurrence. Recall that we analyse causation between events relative to a causal model $\langle M, V \rangle$, where the causal model represents the causal scenario under consideration. We thus need a correspondence between neuron diagrams and causal models.

Here is a recipe to translate an arbitrary neuron diagram, as detailed above, into a causal model. Given a neuron diagram, the corresponding causal model can be constructed in a stepwise fashion. For each neuron n of the neuron diagram,

(i) assign n a propositional variable A.

(ii) If n fires, add the positive literal A to the set V of literals.

(iii) If n does not fire, add the negative literal $\neg A$ to V.

[1] This being quoted, there are some shortcomings of neuron diagrams. For details, see Hitchcock (2007).

(iv) If n has an incoming arrow, write on the right-hand side of A's structural equation a propositional formula ϕ such that ϕ is true iff n fires.

The catch-all condition (iv) stands for the set of the following clauses. (iv$'$) For each stimulatory arrow ending in a normal neuron n, add disjunctively to the right side of A's structural equation the variable that corresponds to the neuron where the arrow originates. (iv$''$) For each pair of stimulatory arrows ending in a stubborn neuron n, add disjunctively to the right side of A's structural equation the conjunction of the two variables that correspond to the two neurons where the arrows originate. (vi$'''$) For each inhibitory arrow ending in n, add conjunctively to the right side of A's structural equation the negation of the variable that corresponds to the neuron where the arrow originates.

One can thus read off a neuron diagram its corresponding causal model: if a neuron is shaded gray, A is in the set V of literals of the corresponding causal model; if a neuron is white, $\neg A$ is in V. And the pattern of arrows translates into a set of structural equations. The translation scheme shows that there is a principled transition from simple neuron diagrams to our causal models.

We add a feature to neuron diagrams in order to represent a removal of information: dotted nodes. Dotted nodes represent neurons about which there is no information as to whether or not they fire. A neuron is dotted iff V' contains no information as to whether or not the neuron fires. If, for example, $A \notin V$ and $\neg A \notin V$, the corresponding neuron will be dotted.

The classical causal scenarios like many others can be represented by a neuron diagram. But this is not true of all causal scenarios. The latter class of scenarios we will represent by dependency diagrams, which are similar to neuron diagrams but more general. We explain dependency diagrams when we need them in Section 7 of Chapter 3.

2 Conjunctive Causes

Let's recall our analysis from Chapter 2: C is a cause of E relative to a causal model $\langle M, V \rangle$ iff C and E are actual in it, there is an active path from C to E in a causal model $\langle M', V' \rangle$ uninformative on both, and the structural equations of C's descendants are in M'. Candidate cause and putative effect will all be actual in the scenarios to come. To check for causation thus boils down to answering two questions:

(a) Is there a causal model uninformative on C and E such that all structural equations of C's descendants are in M'?

(b) If so, can we in a forward-directed way infer E from C in the uninformative model of (a) such that the inferential network of this deduction is an active path?

C is a cause of E iff both questions are answered by yes. In the following sections, we show how our analysis deals with classic causal scenarios.

In a scenario of conjunctive causes, an effect occurs only if more than one cause obtains. Let's say a tree fell only because its roots were weak and the wind blew. The following neuron diagram depicts such a scenario, in which two causes—the tree's having weak roots and the wind blowing—are necessary for an effect to occur—the tree's falling.

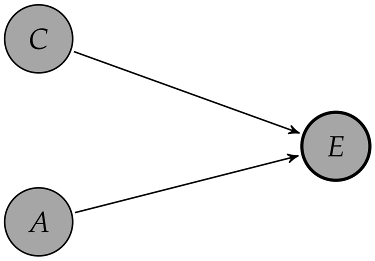


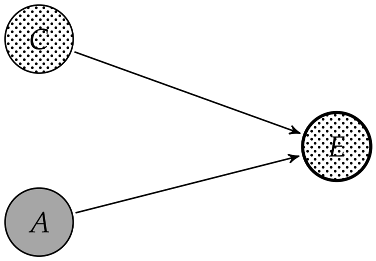
Figure 3: Conjunctive causes

The neurons C and A fire. Together they bring the stubborn neuron E to fire. Stubborn neurons cannot be activated by the firing of a single neuron. Had one of C and A not fired, E would not have been excited. Hence, the firing of both neurons is necessary for E's excitation. Our recipe translates this neuron diagram into the following causal model $\langle M, V\rangle$:

$E = C \wedge A$
C, A, E

The structural equation for E is *conjunctive*: the occurrence of both events, C and A, is necessary for E to occur.

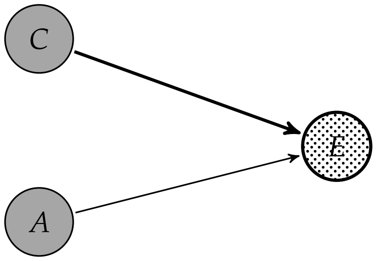
Relative to $\langle M, V\rangle$, C is a cause of E. C and E occur in the original causal model. Furthermore, there is the following causal model $\langle M, V'\rangle$ which is uninformative on C and E:



$E = C \wedge A$
A

Figure 4: Agnostic model for conjunctive causes

We can infer E from $\langle M, \emptyset\rangle[V'][C]$ rather directly using the equation $E = C \wedge A$. The inferential network of this deduction is $C \rightarrow E$. This is a sequence, and therefore an active path by Proposition 4. The crucial point is that each inference to a literal—in the deduction of the effect E—depends directly or indirectly on the candidate cause C. By Proposition 4 we know that any inferential network which has the form of a sequence is an active path. The following figure visually displays the active path by a thick arrow:



$E = C \wedge A$
C, A

Figure 5: Active path from C to E

We have shown that C is a cause of E. Due to the symmetry of the scenario, A is a cause of E as well, as common sense has it.

3 Overdetermination

In a scenario of overdetermination, an effect is overdetermined by more than one event. An example runs as follows: a prisoner is shot by two soldiers at the same time, and each of the bullets is fatal without any temporal precedence. Arguably, each of the shots should qualify as a cause of the death of the prisoner. The following neuron diagram depicts such a scenario, in which an effect is overdetermined by two causes:

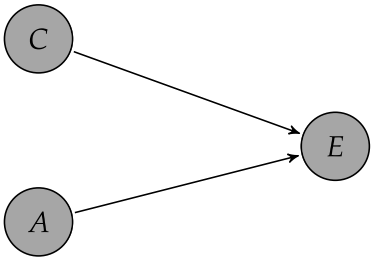


Figure 6: Overdetermination

Neuron C and neuron A fire. The firing of each of C and A alone suffices to excite neuron E. The common firing of C and A overdetermines E to fire. Arguably, the firing of C is a cause of E's excitation, and so is the firing of A. Our recipe translates this neuron diagram into the following causal model:

$E = C \vee A$
C, A, E

The scenario of overdetermination differs from the scenario of conjunctive causes only in the structural equation for E. While the structural equation is *conjunctive* in the latter scenario, the equation is *disjunctive* in the overdetermination scenario. The occurrence of one of the events, C or A, is sufficient for E to occur.

Relative to $\langle M, V \rangle$, C is a cause of E. There is the following causal model $\langle M, V' \rangle$ which is uninformative on C and E:

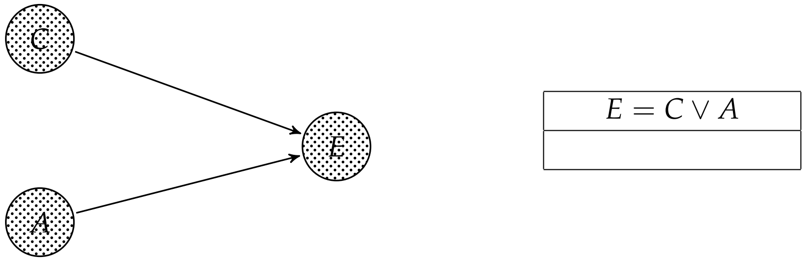


Figure 7: Agnostic model for overdetermination

We can infer E from $\langle M, \emptyset \rangle [V'][C]$ such that the inferential network of this deduction is $C \rightarrow E$. This is a sequence, and so an active path by Proposition 4. Obviously, each inference to a literal depends on the candidate cause C. Here is a visualization of the deduction and the active path:

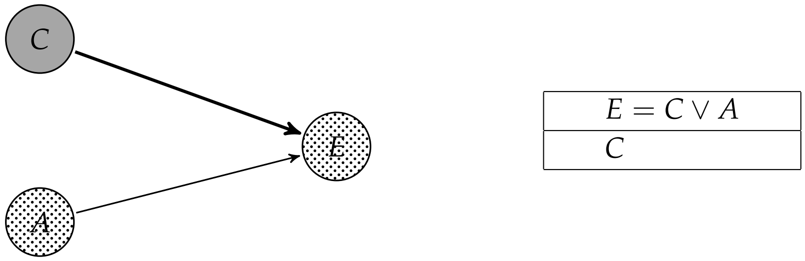


Figure 8: Active path from C to E

We have shown that C is a cause of E. Due to the symmetry of the scenario, A is also a cause of E, as it should be.

4 Early Preemption

In a preemption scenario, an effect E is caused by a genuine cause C. But even if C had not occurred, E would have been brought about by a backup event A. As it is, however, C caused E and A did not. An example of early preemption runs as follows. Suzy (C) and Billy (A) each throw a rock at a window. Suzy's rock deflects Billy's mid-flight so that Billy's does not touch the window ($\neg B$). Only Suzy's rock impacts upon the window (D) and it shatters (E). Had Suzy not thrown, however, Billy's rock would not have been deflected and would have shattered the window. Paul and Hall (2013, p. 75) take the following neuron diagram as canonical for the structure of early preemption:

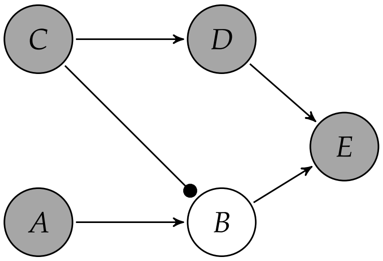
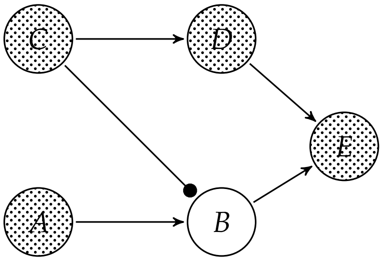


Figure 1: Early preemption

C's firing excites neuron D, which in turn leads to an excitation of neuron E. At the same time, C's firing inhibits the excitation of B. Had C not fired, however, A would have excited B, which in turn would have led to an excitation of E. The actual cause C preempts the mere would-be cause A. Our recipe translates this neuron diagram into the following causal model $\langle M, V \rangle$:

$$\begin{array}{|l|}\hline D = C \\ B = A \wedge \neg C \\ E = D \vee B \\ \hline C, A, D, \neg B, E \\ \hline \end{array}$$

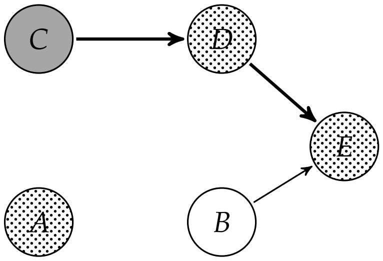
Relative to the model $\langle M, V \rangle$, C is a cause of E. There is the following causal model $\langle M, V' \rangle$ which is uninformative on C and E:



$$\begin{array}{|l|}\hline D = C \\ B = A \wedge \neg C \\ E = D \vee B \\ \hline \neg B \\ \hline \end{array}$$

Figure 9: Agnostic model for early preemption

We can infer E from $\langle M, V' \rangle [V'][C]$ such that the inferential network of this deduction is $C \rightarrow D \rightarrow E$. This sequence is an active path by Proposition 4. Each inference to a literal depends on the candidate cause C, at least indirectly. Here is a visualization of the active path:



$D = C$ $E = D \vee B$
$C, \neg B$

Figure 10: Active path from C to E

We have shown that C is a cause of E.

It remains to show that A is not a cause of E relative to $\langle M, V \rangle$. And indeed, there is no agnostic model $\langle M', V' \rangle$ for which there is a deduction of E from $\langle M', V' \rangle [V'][A]$ such that the inferential network of this deduction is an active path. The demonstration of this claim is a bit complex, though. There are six causal models $\langle M', V' \rangle$ which are uninformative on the candidate cause A and the effect E, given the constraint that M' contains the structural equations of A's descendants: $\langle M, \emptyset \rangle$, $\langle M, \{\neg B\} \rangle$, $\langle M_D, \emptyset \rangle$, $\langle M_D, \{\neg B\} \rangle$, $\langle M_D, \{C\} \rangle$, and $\langle M_D, \{C, \neg B\} \rangle$. The latter four causal models result from suspending judgement concerning the structural equation of D. M_D is obtained from M by removing this equation.

Let us first consider the causal models $\langle M, \{\neg B\} \rangle$, $\langle M_D, \{\neg B\} \rangle$, and $\langle M_D, \{C, \neg B\} \rangle$. Notice that intervening by $\{\neg B\}$ removes the structural equation of B. As a result, there is no deduction of E from these causal models once we have intervened by $\neg B$ and A. By contrast, intervening by $\emptyset$ leaves the structural equations of the causal model unaltered. In more formal terms, it holds that $\langle M, \emptyset \rangle [\emptyset] = \langle M, \emptyset \rangle$. Let us therefore move on to the agnostic model $\langle M, \emptyset \rangle$, and study the deductions of E from this model.

We can infer the effect E from $\langle M, \emptyset \rangle [\emptyset][A]$=$\langle M, \emptyset \rangle [A]$. Reasoning by cases is needed for this inference. Recall the deduction from the previous chapter: if we assume $\neg C$, we can infer B from A by the structural equation $B = A \wedge \neg C$. And from B we can infer E. If we assume C, we can infer E

via D. Hence, whatever assumption we make about the value of C, E can be inferred. The inferential network looks as follows:

$$A \rightarrow B \rightarrow E' \rightarrow E \leftarrow E'' \leftarrow D.$$

This inferential network is not an active path: D and E'' do not depend on the candidate cause A.

Moreover, we can do reasoning by cases with respect to B. If B is true, we can infer E directly. If B is not true, we can infer C from A and $\neg B$ by an indirect proof. From C we can infer E via D. The inferential network is as follows:

$$E' \rightarrow E \leftarrow E'' \leftarrow D \leftarrow C \leftarrow A.$$

This inferential network is not an active path: E' does not depend on the candidate cause A. Note that assumptions of subproofs are left out of the inferential network by definition. This is why the literals B and $\neg B$ have no occurrences in the network.

Yet another option is to do reasoning by cases with respect to D. The result is the same. We have two subproofs concluding with E, but the network of the complete deduction is not an active path. Both subproofs violate the condition that all inferred literals depend on A. This is particularly obvious for the inferential step from D to E, which is at the beginning of the subproof on assumption of D.

Finally, we must wonder if E could be derived by a proof by contradiction in such a manner that the network of this deduction is an active path. Suppose, for contradiction, that E is false. Since $\neg E$ is an assumption of a subproof and not an explicit premise, we can draw backward-directed inferences from $\neg E$. Specifically, we can infer $\neg B$ and $\neg D$ using the equation $E = B \vee D$. For these inferences, we need another subproof which starts with $B \vee D$. By the equation $E = B \vee D$, the disjunction $B \vee D$ gives us E, which contradicts $\neg E$. Thus we have inferred $\neg(B \vee D)$, from which we can infer $\neg B$ and $\neg D$ by classical reasoning. Before completing this indirect proof, we can already tell that it does not lead to a deduction which has an active path. None of the inferential steps depends on the candidate cause. The assumption of A did not even come into play yet.

It may come as a surprise that this fragment of an indirect proof contains a backward-directed section: we inferred $\neg B$ and $\neg D$ from $\neg E$ in the context of M. Notice, for clarification, that a deduction of E from $\langle M, V\rangle[V][C]$ may contain backward-directed sections of reasoning in a subproof. The presence of such sections in a subproof does not contradict our claim that all consequences of $\langle M, V\rangle[V][C]$ are forward-directed. Likewise for all conclusions in the main proof of a deduction.

It remains to consider the possibility of a deduction of E from the agnostic models $\langle M_D, \emptyset\rangle[A]$ and $\langle M_D, \{C\}\rangle$. However, there is no such deduction. Reasoning by cases with regard to the variable C does not work here because M_D lacks the structural equation of D. Likewise, reasoning by cases with regard to D and B fails to work. Finally, it is not possible to derive a contradiction from the assumption $\neg E$ in the context of the causal models $\langle M_D, \emptyset\rangle[A]$ and $\langle M_D, \{C\}\rangle[A]$.

We have now studied four different deductions of the effect E from the causal model $\langle M, \emptyset\rangle[\emptyset][A]$. None of these deductions has an active path. The other agnostic models lack a deduction of E altogether. $\langle M, V\rangle[A] \vdash E$ fails to hold for these models. Have we thereby established that the preempted cause A is not genuine on our analysis? There remains one problem: how can we show that we have not overlooked a deduction which has the desired property that all inferred literals depend on the candidate cause A? Perhaps, there is some combination of a direct deduction with an indirect one which satisfies our definition of an active path. In the next section, we will show that we can dispense with indirect proofs when searching for a deduction of the effect with an active path. This result enables us to show in a conclusive manner hat a preempted cause does not count as genuine on our analysis.

5 Direct Deductions and Genuine Causation

Let us understand the notion of direct deduction as antonym to the notion of indirect proof. A direct deduction may contain reasoning by cases, but does not contain any indirect proof. Let us call a deduction *strictly direct* iff it contains no subproofs at all. We have seen how an indirect proof works for causal models. To prove ϕ, we assume $\neg\phi$. Then we derive a contradiction from $\neg\phi$ in the context of the premises and equations of the causal model. From this contradiction we infer ϕ. Indirect proofs are also referred to as *proofs by contradiction.*

What is wrong with indirect proofs when trying to show that there is an active path from the candidate cause to the effect? In the preemption scenario, all deductions with an indirect proof violate either the condition of non-redundancy or the condition that all inferred literals depend on the candidate cause. This problem arises for deductions from both the genuine and the preempted cause. We can show that indirect proofs, if used in a deduction with an active path, are dispensable: if there is a deduction which has an active path, then there is a direct deduction with an active path.

Proposition 5. Let $\langle M, V \rangle$ be a causal model which is uninformative on C and E. Suppose there is a deduction of E from $\langle M, V \rangle [V][C]$ such that this deduction has an active path. Then there is such a deduction which is direct with respect to all causal inferences.

Only indirect proofs with causal inferences are dispensable. Other types of indirect proofs are not. A case in point is the deduction of $\neg(A \wedge B)$ from $\neg A$. This deduction, however, does not have a causal meaning. We need an indirect proof for it only if we work with fully formal deductions. Recall from the previous chapter that an inference is said to be causal iff it uses a structural equation directly or indirectly in a subproof.

Proposition 5 helps us simplify the search for a deduction which has an active path. We are now in a position to show in a conclusive manner that a preempted cause is not genuine on our analysis. Let us resume the discussion of the preemption scenario from the previous section. Recall the causal models which are uninformative on the preempted cause A and the effect E: $\langle M, \emptyset \rangle$, $\langle M, \{\neg B\} \rangle$, $\langle M_D, \emptyset \rangle$, $\langle M_D, \{\neg B\} \rangle$, $\langle M_D, \{C\} \rangle$, and $\langle M_D, \{C, \neg B\} \rangle$. Further, recall that M_D is obtained from M by removing the structural equation of D.

Suppose the preempted cause A were to count as genuine on our analysis. By Proposition 5, this implies that there is a direct deduction of E from $\langle M, V \rangle [V][A]$ for at least one of the agnostic models. However, there is no such deduction. For this to be seen, let us first consider the agnostic models $\langle M, \{\neg B\} \rangle$, $\langle M_D, \{\neg B\} \rangle$, and $\langle M_D, \{C, \neg B\} \rangle$. As explained above, intervention by $\neg B$ removes the structural equation of B. Hence, there is no directed path from A to E in the causal graphs of said causal models after intervention by $\neg B$ and A. By Proposition 2, this implies that there is no deduction of E from any of these causal models.

Likewise, there is no deduction of the effect E from the preempted cause A for the agnostic models $\langle M_D, \{C\} \rangle$ and $\langle M_D, \emptyset \rangle$ in the first place. We can show this by a combination of deductive reasoning and semantic considerations. Let us suppose $\neg E, C$, and A. From C and A we can infer $\neg B$. From $\neg E$ we can infer $\neg D$. Thus we obtain a complete valuation $V'' = \{A, C, \neg B, \neg D, \neg E\}$. Notably, this valuation satisfies all structural equations in M_D and it satisfies C. Hence, it is consistent to assume that the effect E does not occur for both of the agnostic models $\langle M_D, \{C\}\rangle$ and $\langle M_D, \emptyset\rangle$. Hence, there is no deduction of E from A for these models.

Finally, it remains to consider the agnostic model $\langle M, \emptyset\rangle$. Here we have a deduction of E from $\langle M, \emptyset\rangle[A]$. However, reasoning by cases is needed for this deduction. At least one subproof contains an inferential step which does not depend on the candidate cause A. More specifically, for each deduction of E using a proof by cases, a subproof is needed whose first inferential step toward E is independent of A. We have seen this problem to arise for proofs by cases with regard to C, D, and B in the previous section. By Proposition 5, we can ignore proofs by cases with regard to E since the subproof which starts with $\neg E$ amounts to an indirect proof of E.

Thus we have shown that there is no deduction of E from $\langle M, \emptyset\rangle[A]$ which has an active path. Since there is no deduction of E from A for the other agnostic models, this implies that there is no deduction of E from $\langle M, V'\rangle[V'][A]$ with an active path for any of the agnostic models. The preempted cause A does therefore not count as genuine on our analysis.

6 Late Preemption

Lewis (1986b, p. 200) subdivides preemption into early and late. We have discussed early preemption in the previous sections: a backup process is cut off before the process started by the preempting cause could bring about the effect. In scenarios of late preemption, by contrast, the backup process is cut off only because the genuine cause brings about the effect before the preempted cause could do so. Lewis (2000, p. 184) provides the following story for late preemption:

> Billy and Suzy throw rocks at a bottle. Suzy throws first, or maybe she throws harder. Her rock arrives first. The bottle shatters. When Billy's rock gets to where the bottle used to be, there is nothing there but flying shards of glass. Without Suzy's

> throw, the impact of Billy's rock on the intact bottle would have been one of the final steps in the causal chain from Billy's throw to the shattering of the bottle. But, thanks to Suzy's preempting throw, that impact never happens.

Crucially, the backup process initiated by Billy's throw is cut off only by Suzy's rock impacting the bottle. Until her rock impacts the bottle, there is always a backup process that would bring about the shattering of the bottle an instant later.

How to best represent late preemption in neuron diagrams and causal models is somewhat controversial (Hall 2007, Hitchcock 2007, Paul and Hall 2013). We follow Halpern and Pearl (2005, pp. 861–2) who propose a causal model for late preemption corresponding to the following neuron diagram.

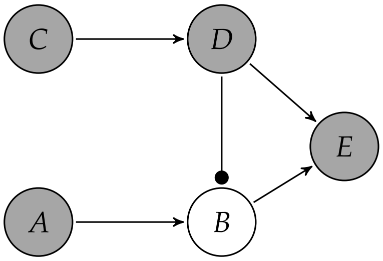


Figure 11: Late preemption

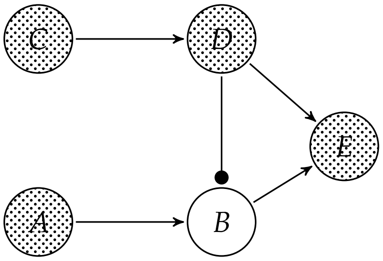
Suzy throws her rock (C) and Billy his (A). Suzy's rock impacts the bottle (D), and so the bottle shatters (E). Suzy's rock impacting the bottle (D) prevents Billy's rock from impacting the bottle ($\neg B$).

Our recipe translates the neuron diagram of late preemption into the following causal model $\langle M, V \rangle$:

$D = C$ $B = A \wedge \neg D$ $E = D \vee B$
$C, A, D, \neg B, E$

Only the equation for B differs from the causal model of early preemption: the occurrence of B requires A to occur and the absence of D instead of the absence of C. This difference seems negligible given that D occurs just in case C occurs. It is thus unsurprising that our analysis treats late preemption analogous to early preemption.

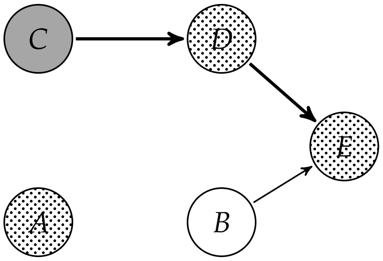
Relative to $\langle M, V\rangle$, C is a cause of E. There is the following causal model $\langle M, V'\rangle$ which is uninformative on C and E:



$D = C$ $B = A \wedge \neg D$ $E = D \vee B$
$\neg B$

Figure 12: Agnostic model for late preemption

We can infer E from $\langle M, \emptyset\rangle[V'][C]$ such that the inferential network of this deduction is $C \rightarrow D \rightarrow E$. The network is an active path since all inferences to a literal depend on the candidate cause C. The active path may be graphically depicted as follows:



$D = C$ $E = D \vee B$
$C, \neg B$

Figure 13: Active path from C to E

We have shown that C is a cause of E.

A is not a cause of E relative to $\langle M, V\rangle$. The reasoning is analogous to the one of early preemption in the previous section.

7 Trumping Preemption

Our analysis solves preemption without intermediate variables. For this to be seen, let us consider a scenario known as *trumping*, which is particularly troublesome for counterfactual theories of causation. Schaffer (2000) provides an example akin to the following. It is a rule of the military that commands from higher-ranking officers trump those of lower rank. The major and the sergeant stand before a soldier, both shout 'Advance!' at the same time, and the soldier advances. It is the major's command, and not the sergeant's, that causes the soldier to advance. The sergeant's command has not been effective because it has been trumped by the major's.

There is some controversy about how to best represent a scenario of trumping. This is clear: the soldier's advancing is redundantly caused. The soldier would still have advanced if the sergeant had shouted 'Advance!' and the major had been silent, or if the major had shouted 'Advance!' and the sergeant had been silent. And the redundant causation is asymmetrical. The major's command is causally efficient, while the sergeant's is not.

Schaffer (2000) suggested to view the scenario of trumping as a case of preemption. This makes perfect sense. Recall that we have characterized preemption through the notion of a backup cause. The preempted cause is a backup cause in that it would be effective if the genuine cause were to be absent. Put very briefly, the preempted cause acts if the genuine cause does not. Since, however, the genuine cause is active by assumption in the causal scenario, the preempted cause remains a mere backup, which does not get to act.

Furthermore, we suggest to characterize trumping as a case of *simultaneous preemption*. It is simultaneous in two respects. First, the genuine and the preempted cause occur at the same time. Second, the genuine cause is effective at exactly the same time the preempted cause would be if the genuine cause was not present. Unlike in early preemption, there is no interference by an intermediate event through which the genuine and the preempted cause interact. Unlike in late preemption, the reason why the genuine cause is active is not that it acts faster than the preempted cause. Simultaneous preemption is different from both early and late preemption (Lewis 2000).

The formal representation of trumping is controversial. Unlike in scenarios of early and late preemption, there are no obvious intermediate events—intermediate between the two candidate causes and their effect—which could help us explain why the preempted cause doesn't get to act. A minimalist representation of Schaffer's trumping scenario without intermediate variables goes as follows:

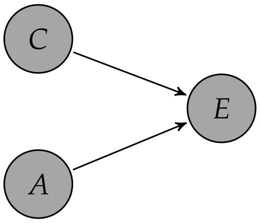


Figure 14: Trumping

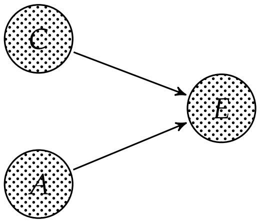
The major's command to advance (C) is sufficient for the advancement of the soldier (E), and so is the command of the sergeant (A) in the absence of higher-ranking commands. In the scenario under consideration, the major actually gives the soldier a command to advance, and so does the sergeant.

The structural equation of the trumping scenario is $E = C \vee (A \wedge \neg C)$: the soldier advances just in case the major gives the command to advance, or the sergeant does and the major does not. Note that the structural equation contains strictly more information than the graphical representation, which looks just like a scenario of overdetermination. Indeed, the graphical representation is a dependency diagram. Such a diagram represents only the dependencies between variables and what events occur. An occurring event is represented by a grey node, an absence by a white node. Unlike a neuron diagram, a dependency diagram does not allow us to read off the structural equations of the causal scenario. Directed arrows, in particular, stand for any type of dependence. A dependency diagram therefore contains less information than a neuron diagram.

Here is the resulting causal model for the trumping scenario:

$E = C \vee (A \wedge \neg C)$
C, A, E

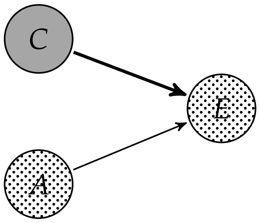
Our analysis delivers the intuitive verdicts for trumping, even without intermediate variables or other extensions. The demonstration is straightforward. Relative to $\langle M, V \rangle$, C is a cause of E. There is the following causal model $\langle M, V' \rangle$ which is uninformative on C and E:



$E = C \lor (A \land \neg C)$

Figure 15: Agnostic model for trumping

We can infer E from $\langle M, \emptyset \rangle [\emptyset][C]$ such that the network of this deduction is $C \rightarrow E$. Obviously, each inference to a literal depends on C, and so the network is an active path. Here is a visualization:



$E = C \lor (A \land \neg C)$
C

Figure 16: Active path from C to E

We have shown that C is a cause of E.

It remains to show that A is not a cause of E. There is only one causal model which keeps the structural equation and is uninformative on A and E: $\langle M, \emptyset \rangle$. We can infer E from $\langle M, \emptyset \rangle [V'][A]$, but the inferential network of this deduction is not an active path:

$$A \rightarrow E' \rightarrow E \leftarrow E''.$$

The inferential network stands for a deduction with reasoning by cases. If C is not actual, E can be inferred by A. If C is actual, E can be inferred therefrom without A. This means the inferential step does not depend on A. There is no other direct deduction of E from $\langle M, \emptyset \rangle [\emptyset][A]$ which has an active path. There is also no other direct deduction of E from any causal model $\langle \emptyset, V' \rangle [V'][A]$ which is uninformative on E. By Proposition 5, we know that we can ignore indirect deductions of E. We have thus shown that A is not a cause of E.

We find it quite remarkable that our analysis can discriminate between genuine and preempted causes on the minimalist causal model of trumping. No intermediate variables are needed. We are not aware of any other causal model analysis which delivers this result. There is a deeper reason why analyses of actual causation using the standard account of causal models have difficulties making the discrimination in question.

Note that the right-hand side of the structural equation $E = C \vee (A \wedge \neg C)$ is logically equivalent to the right-hand side of $E = C \vee A$. This implies that the two equations are indistinguishable if understood in a purely semantic way. Hence, a purely semantic account of structural equations, such as the one in Halpern (2000) and Halpern and Pearl (2005), is unable to distinguish between genuine and trumped causes in the trumping scenario of this section from the outset. To cope with this trumping scenario, accounts relying on purely semantic structural equations seem to require a hyperintensional refinement—a refinement which allows to distinguish between logical equivalents (Berto and Nolan 2023).

8 Prevention

In a prevention scenario, an event prevents another from occurring. However, had the event not occurred, the other would not have been prevented, and so would have occurred. Here is an example. An assassin poisons victim's coffee (D). Victim's bodyguard puts in an antidote (C), which prevents the poison from killing victim ($\neg E$). Had bodyguard not put in the antidote, victim would have died. Paul and Hall (2013, p. 174) represent the basic scenario of prevention by the following neuron diagram:

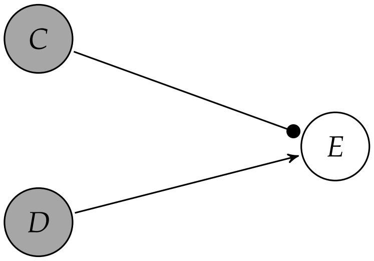
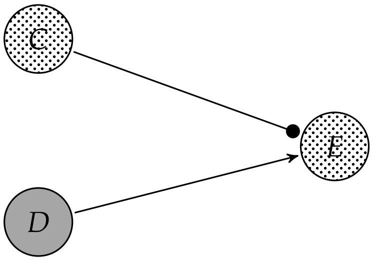


Figure 17: Prevention

Neuron C fires and thereby inhibits that neuron E fires. E would have been excited by D if the inhibitory signal from C had been absent. But as it is, C prevents E from firing. That is, C causes $\neg E$ by prevention. Our recipe translates the neuron diagram of prevention into the following causal model $\langle M, V \rangle$:

$E = \neg C \wedge D$
$C, D, \neg E$

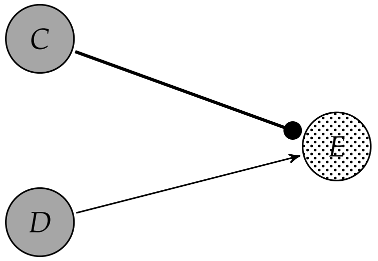
Relative to $\langle M, V \rangle$, C is a cause of $\neg E$. There is the following causal model $\langle M, V' \rangle$ which is uninformative on C and $\neg E$:



$E = \neg C \wedge D$
D

Figure 18: Agnostic model for prevention

We can infer $\neg E$ from $\langle M, \emptyset \rangle [V'][C]$ such that the inferential network of this deduction is an active path: $C \rightarrow \neg E$. Obviously, each inference to a literal in this network depends on the candidate cause C. We display the active path by the following figure:



$E = \neg C \wedge D$
C, D

Figure 19: Active path from C to E

We have shown that C is a cause of $\neg E$.

Finally, D is not a cause of $\neg E$ relative to $\langle M, V\rangle$. Note that any causal model $\langle M, V'\rangle$ which is uninformative on D and $\neg E$ is also uninformative on C. Because of condition (3) in the definition of $\gg$, we need to retain the structural equation of E when looking for an uninformative causal model with an active path from D to $\neg E$. Hence, there is no causal model $\langle M', V'\rangle$ which is uninformative on D and $\neg E$ such that $\neg E$ can be inferred from $\langle M', V'\rangle[V'][D]$. And so there cannot be an uninformative model $\langle M, V'\rangle$ with an active path from D to $\neg E$.

9 Double Prevention

In a scenario of double prevention, an event prevents a threat for another event's occurrence. More precisely, an event prevents an event which—had it occurred—would have prevented a third event. We say an event C double prevents an event E if C prevents an event that—had it occurred—would have prevented E. Paul and Hall (2013, p. 175) provide an example:

> David makes the coffee (A), and fills his cup (E). Meanwhile, Steffi scoops up James the cat (C) as he lashes his tail wildly (B); her quick action prevents a disastrous spilling (D). Most of us want to say that Steffi saved the coffee—that is, C was among the causes of E.

They represent the structure of double prevention by the following neuron diagram:

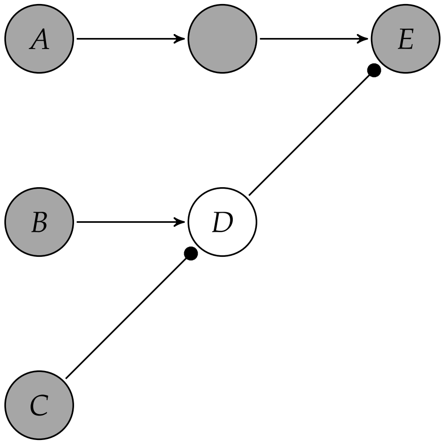


Figure 20: Double prevention

The characteristic structure of double prevention is this: *C*'s firing prevents *D*'s firing, which would have prevented *E*'s firing. This structure exhibits a counterfactual dependence: given that *B* fires, *E*'s firing counterfactually depends on *C*'s firing. If *C* had not fired, *D* would fire, and thereby prevent *E* from firing. *C*'s firing prevents a threat for *E*'s firing, namely the threat originating from *B*'s firing. In short, *C*'s firing double prevents *E*'s firing.

C is arguably a cause of *E* in the present scenario of double prevention: Steffi's scooping up James the cat prevents the spilling of the coffee, which would have prevented the filling of the cup. Our recipe translates the neuron diagram of Figure 20 into the following causal model $\langle M, V \rangle$:

$D = B \wedge \neg C$ $E = A \wedge \neg D$
$A, B, C, \neg D, E$

Relative to $\langle M, V \rangle$, *C* is a cause of *E*. There is the following causal model $\langle M, V' \rangle$ which is uninformative on *C* and *E*:

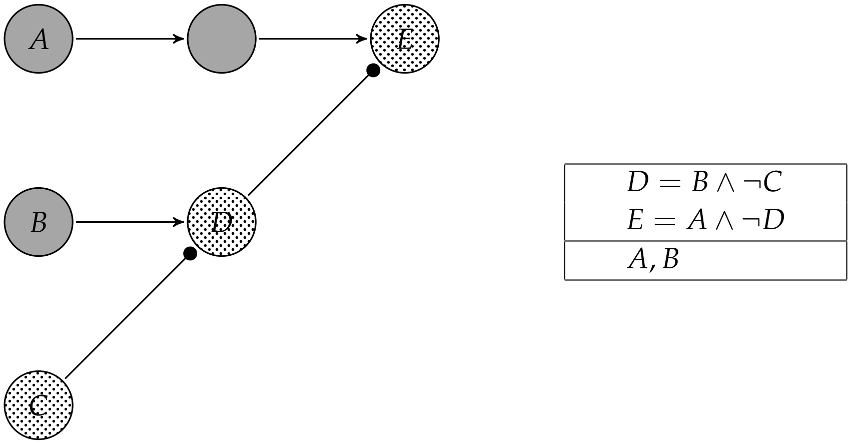


Figure 21: Agnostic model for double prevention

We can infer E from $\langle M, \emptyset \rangle [V'][C]$ such that the inferential network of this deduction is $C \rightarrow \neg D \rightarrow E$. This is a sequence, and therefore the network is an active path. Each inference to a literal depends on the candidate cause C. Here is a graphical representation:

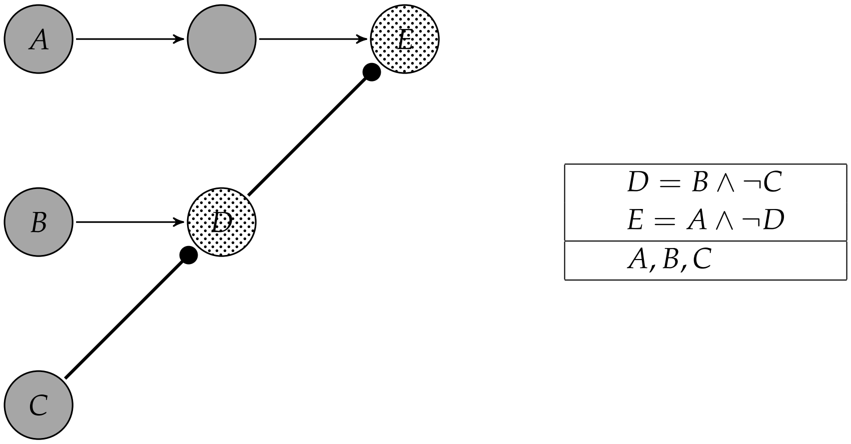


Figure 22: Active path from C to E

We have shown that C is a cause of E.

Chapter 4

Non-Transitivity

In this chapter, we show that our analysis captures our judgments in scenarios which suggest that causation is not transitive. In particular, we show that our analysis captures our judgments in scenarios known as short circuit, extended double prevention, and some switching scenarios.

Several scenarios have been put forth which suggest that our causal judgments are not transitive (McDermott 1995, Lewis 2000, Paul 2000). The transitivity of causation means this: whenever C is a cause of A and A is a cause of E, then C is a cause of E. It seems often plausible to judge C a cause of E if you judge C a cause of A and A a cause of E. In light of our analysis, the plausibility is not hard to explain: if there is an active path from C to A in an uninformative model and there is one from A to E in the same uninformative model, then there is one from C to E in this uninformative model. We suggest this is why it is often appropriate to judge that C causes E by tracing a causal chain from C over other events to E. Our analysis can explain why transitivity is plausible for causation. And yet it does not rely on transitivity to handle certain causal scenarios. We are thus free to deny that our causal judgment is invariably transitive.

The challenge of non-transitivity is to show that certain events are non-causes—even though they share certain structural properties with genuine causes. In this respect, the challenge resembles the preemption problem: a preempted cause does not count as genuine—even though we can infer from it the effect in an agnostic model. We have solved the latter problem by requiring that each inferential step to an event or absence must depend on the candidate cause in the inferential reconstruction of a causal process.

Our solution to the challenge of non-transitivity exploits a subtle condition in our definition of an epochetic conditional for causal models. To show that $C \gg E$ holds, it is not enough to find an active path from C to E in a causal model which is agnostic about C and E. The active path must be found for an agnostic causal model which contains the structural equations of all descendants of the candidate cause C. The rational for this condition is to leave intact the inferential relations between the candidate cause and its potential effects. We will show that there is no such model for a number of causal scenarios where our causal judgements are not transitive.

1 Short Circuit

One of the examples against transitivity goes as follows. A boulder is dislodged and rolls toward a hiker. The hiker sees the boulder coming and ducks, so that she does not get hit by the boulder. If the hiker had not ducked, however, the boulder would have hit her (Hitchcock 2001, cf. p. 276).

The boulder scenario seems to show that there are cases where causation is not transitive: the dislodged boulder causes the ducking of the hiker, which in turn causes the hiker to remain unscathed. But it is counterintuitive to say that the dislodging of the boulder causes the hiker to remain unscathed. The structure of the boulder scenario can be represented by the following neuron diagram:

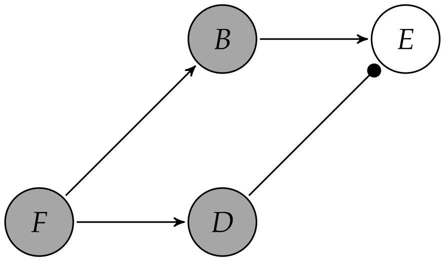


Figure 23: Short circuit

Hall (2007, p. 36) calls the network of Figure 23 a *short circuit*: the boulder's dislodgement (F) threatens to hit the hiker by a rolling boulder (B), and at the same time provokes an action—the ducking (D)—that prevents this threat from being effective ($\neg E$).

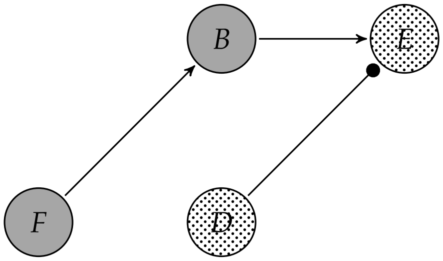
In neuron speak, F fires and thereby excites neuron B to fire, which in turn threatens to excite neuron E. At the same time, F's firing excites neuron D, whose firing prevents E from firing. So F's firing creates a process via B that threatens to bring about E and at the same time initiates another process via D which prevents the threat from being effective. F cancels its *own* threat—the threat via B—to prevent E.

F should not count as a cause of $\neg E$ because F creates *and* cancels the threat to bring about E (Paul and Hall 2013, p. 216). Our recipe translates the neuron diagram of the boulder scenario into the following causal model $\langle M, V \rangle$:

$$\begin{array}{|l|}\hline B = F \\ D = F \\ E = B \land \neg D \\ \hline F, B, D, \neg E \\ \hline\end{array}$$

Relative to $\langle M, V \rangle$, F is not a cause of $\neg E$. For this to be seen, observe that all variables are descendants of F, except for F itself. By condition (3) of the definition of $\gg$, this implies that we cannot suspend judgement on a structural equation in M. Reasoning by cases furthermore reveals that the structural equations in M entail $\neg E$. We can infer $\neg E$ from both $\{F\} \cup M$ and $\{\neg F\} \cup M$, which implies $\langle M, \emptyset \rangle \models \neg E$. Hence, there is no causal model $\langle M', V' \rangle$ wich is uninformative on E and which satisfies condition (3). And so $F \gg \neg E$ fails to hold.

By contrast, D is a cause of $\neg E$. For this to be seen, observe that D is not a descendant of itself. Hence, the structural equation $D = F$ can be removed. There is the following causal model $\langle M', V' \rangle$ which is uninformative on D and $\neg E$:

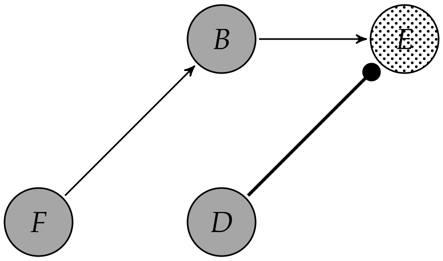


$B = F$
$E = B \wedge \neg D$

F, B

Figure 24: Causal model agnostic on D and E

We can infer $\neg E$ from $\langle M, V' \rangle [V'][D]$ such that the inferential network of this deduction is $D \rightarrow \neg E$. This is an active path since each inference to a literal depends on the candidate cause D. The deduction is displayed by the following figure:



$B = F$
$E = B \wedge \neg D$

F, B, D

Figure 25: Active path from D to $\neg E$

We have shown that D is a cause of $\neg E$, as desired.

Relative to $\langle M, V \rangle$, B is not a cause of $\neg E$. For this to be seen, note that D is necessary to infer $\neg E$, and any causal model uninformative on B and $\neg E$ must either be uninformative on D, or else lack the structural equation of E. In both cases, intervening by B does not bring about $\neg E$. And so $F \gg \neg E$ fails to hold.

2 Extended Double Prevention

Hall (2004, p. 247) presents an extension of the scenario of double prevention depicted in Figure 20. The extended double prevention scenario fits the structure of the following neuron diagram:

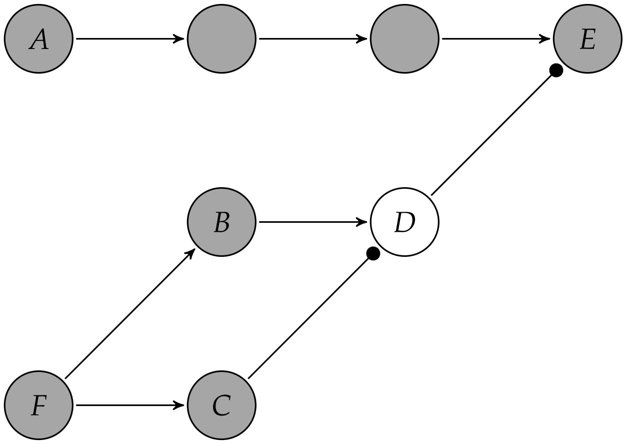


Figure 26: Extended double prevention

Figure 26 extends Figure 20 by neuron F, which figures as a common cause of B and C. The subgraph $F - B - C - D$ is a short circuit. F starts a process via B that threatens to prevent E. At the same time, F initiates another process via C that prevents the threat. F cancels its *own* threat—the threat via B—to prevent E.

In the scenario of double prevention, explained in Section 9 of the previous chapter, the threat for E originated independently of its preventer. Here, by contrast, F creates and cancels the threat to prevent E. This difference is sufficient for F not to be a cause of E (Paul and Hall 2013, p. 216). Observe that the structure characteristic of double prevention is embedded in Figure 26. The firing of neuron C inhibits D's firing that, had it fired, would have inhibited E's firing. Nevertheless, this scenario of extended double prevention exhibits an important difference to its relative of double prevention: E does not counterfactually depend on F. If F had not fired, E would still have fired.

Our recipe translates the neuron diagram of extended double prevention into the following causal model $\langle M, V \rangle$:

$B = F$ $C = F$ $D = B \wedge \neg C$ $E = A \wedge \neg D$
$A, F, B, C, \neg D, E$

Relative to this causal model, F is not a cause of E. For this to be seen, observe that the variables B, C, D, and E are descendants of F. So no structural equation can be removed by condition (3) of the definition of our epochetic conditional $\gg$. Hence, to show that $F \gg E$ holds, we need to find an active path from F to E in an agnostic model $\langle M', V' \rangle$ such that $M' = M$. There is only one such model which is uninformative on F and E:

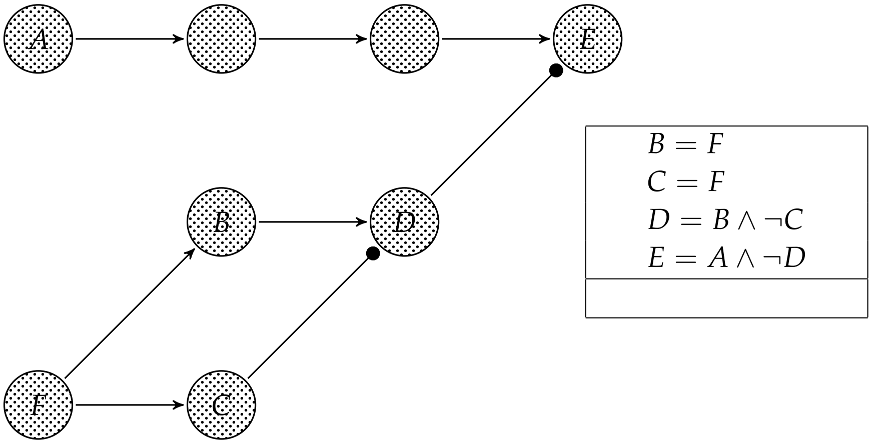


Figure 27: Causal model agnostic on F and E

But we cannot infer E from $\langle M, V' \rangle [V'][F]$. This inference would require that $A \in V'$. But then $\langle M, V' \rangle$ would not be uninformative on E. We have shown that F is not a cause of E, as common sense says.

Relative to $\langle M, V \rangle$, C is a cause of E. There is the following causal model $\langle M', V' \rangle$ which is uninformative on C and E and in which the structural equation $C = F$ has been removed:

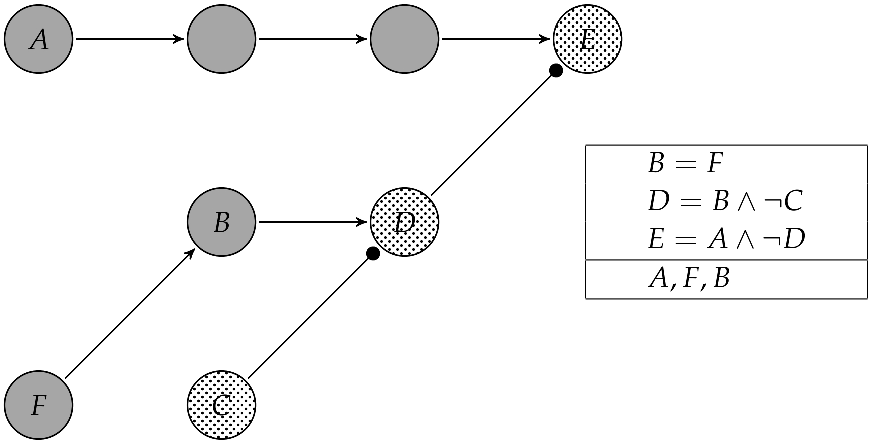


Figure 28: Causal model agnostic on C and E

We can infer E from $\langle M, V' \rangle [V'][C]$ such that the inferential network of this deduction is $C \rightarrow \neg D \rightarrow E$. This is an active path since each inference to a literal depends on the candidate cause C. Here is a visualization of the active path:

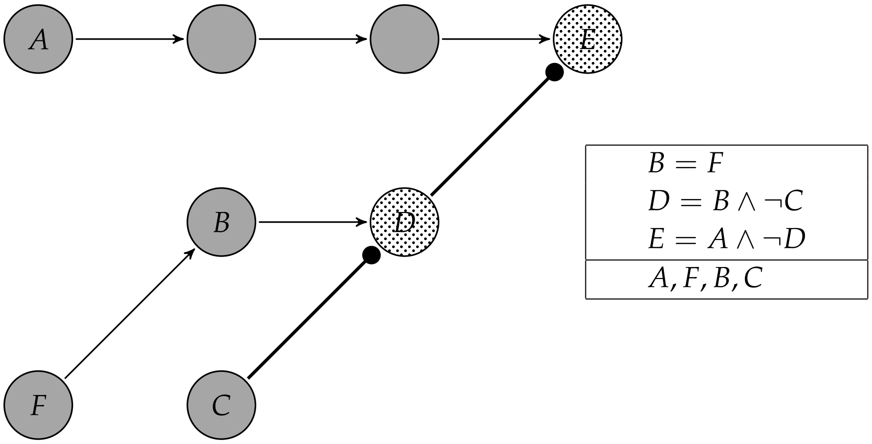


Figure 29: Active path from C to E

We have shown that C is a cause of E.

3 Simple Switch

Switching scenarios are paradigmatic for causal scenarios where our causal judgments are not transitive. In switching scenarios, some event F helps to determine the causal path by which another event is brought about. Crucially, the other event would also occur via an alternative path if F had not occurred. Indeed, if the non-actual switch position $\neg F$ were actual, $\neg F$ would be on this path to E's occurrence. This is a noteworthy difference to preemption scenarios. If Suzy were not to throw her rock, her not throwing would not help to bring about the bottle's shattering.

To make switches more concrete, consider a story provided by Hall (2000, p. 205). Flipper is standing by a switch in the railroad tracks. A train approaches in the distance. She flips the switch (F), so that the train travels down the right track (R), instead of the left ($\neg L$). Since the tracks reconverge up ahead, the train arrives at its destination all the same (E). The commonsensical judgment is that flipping the switch is not a cause of the train's arrival—even though flipping the switch is a cause of the train's travelling on the right track, and the train's travelling on the right track is a cause of the train's arrival (Paul and Hall 2013, p. 232).

We think it is hard to represent switching scenarios by neuron diagrams. One reason is that the two positions of a switch are assumed to be symmetric, while the firing and the non-firing of a neuron are not. A further reason is that a firing 'switch neuron' would activate a neuron and inhibit another. But this is too much: a switch should only determine the path by which an event is brought about. Neuron diagrams introduce an asymmetry with respect to the position of a switch, while there should be none. Hence, we represent the switching scenario by a simple dependency diagram:

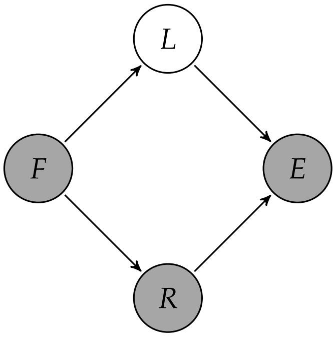


Figure 30: Simple switch

F acts like a switch as to E. And so F does not count as a cause of E, even though F should count as a cause of R and R should count as a cause of E.

The simple switch can be represented by the following causal model $\langle M, V \rangle$:

$$\begin{array}{|c|} \hline L = \neg F \\ R = F \\ E = L \vee R \\ \hline F, \neg L, R, E \\ \hline \end{array}$$

Relative to $\langle M, V \rangle$, F is not a cause of E. For this to be seen, observe that all variables except F are descendants of the candidate cause F. By condition (3) in the definition of $\gg$, this implies that no structural equation can be removed when we look for an agnostic model with an active path. Now, we cannot find an agnostic model with an active path from F to E since there is simply no agnostic model $\langle M', V' \rangle$ such that $M' = M$. Even the causal model $\langle M', \emptyset \rangle$ entails E. For this to be seen, suppose F. (i) From this assumption we can infer E via the literal R. Likewise, (ii) from the assumption of $\neg F$ we can infer E via the literal L. Note that (iii) $F \vee \neg F$ can be derived from the empty premise set in our logic of causal models. (i), (ii), and (iii) imply that $\langle M, \emptyset \rangle$ entails E by soundness of the deductive system. In the absence of a causal model $\langle M, V' \rangle$ which is agnostic on E, $F \gg E$ fails to hold.

Relative to $\langle M, V \rangle$, F is a cause of R. There is, among others, the following causal model $\langle M', V' \rangle$ which is uninformative on F and R:

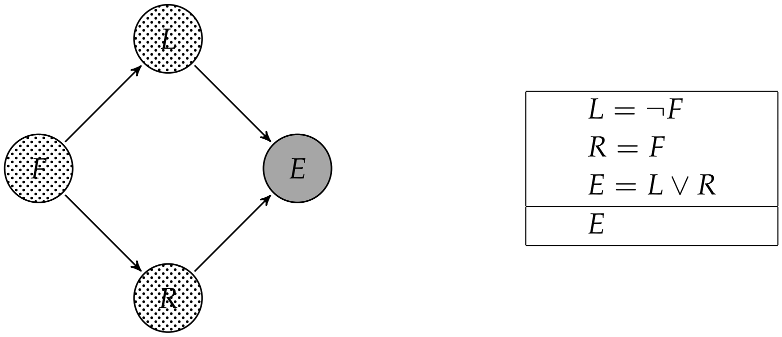


Figure 31: Causal model agnostic on F and R

We can infer R from $\langle M', V'\rangle[V'][F]$ such that the inferential network of this deduction is $F \rightarrow R$. This is an active path since each inference to a literal depends on the candidate cause F. The deduction may be visualized by the following figure:

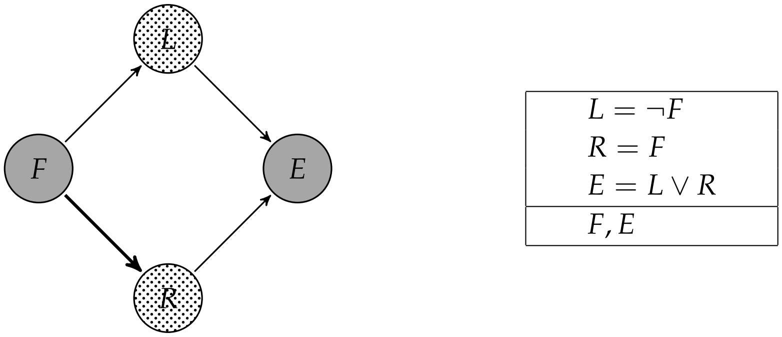
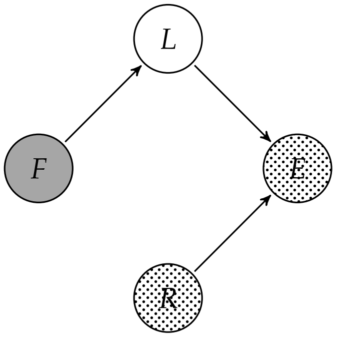


Figure 32: Active path from F to R

We have shown that F is a cause of R, as desired.

Relative to $\langle M, V\rangle$, R is a cause of E. There is, among others, the following causal model $\langle M', V'\rangle$ which is uninformative on R and E:

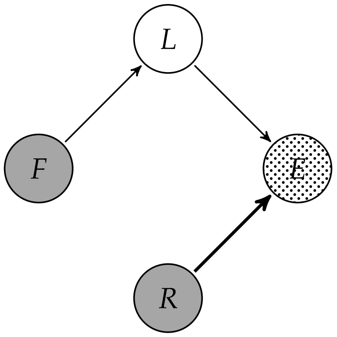


$L = \neg F$
$E = L \lor R$

$F, \neg L$

Figure 33: Causal model agnostic on R and E

We can infer E from $\langle M', V' \rangle [V'][R]$ such that the inferential network of this deduction is $R \rightarrow E$. This is an active path since each inference to a literal depends on the candidate cause R. The active path may be visualized by the following figure:



$L = \neg F$
$E = L \lor R$

$F, \neg L, R$

Figure 34: Active path from R to E

We have shown that R is a cause of E, as common sense has it.

4 Basic Switch

The representation of switching scenarios is somewhat controversial. Paul and Hall (2013, p. 232) represent a 'basic' switch by the following neuron diagram:

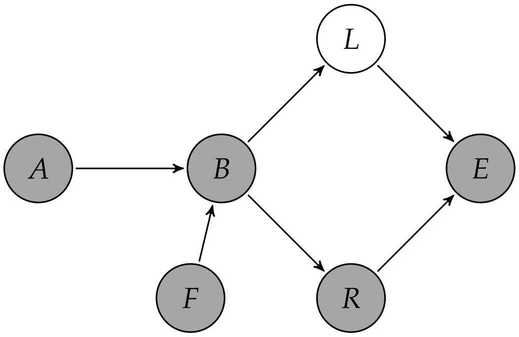


Figure 35: Neuron diagram for basic switch

Paul and Hall 'stipulate' that neuron F acts like a switch as to neuron E. If neuron F fires, the signal from neuron B's firing travels down and excites neuron R. If F does not fire, the signal from B travels up and excites neuron L. Either way, neuron E gets excited and fires.

Paul and Hall's neuron diagram for the basic switch has additional stipulations. As already mentioned, we think it is hard to represent switching scenarios by neuron diagrams. We think the following dependency diagram is better suited to represent the basic switch:

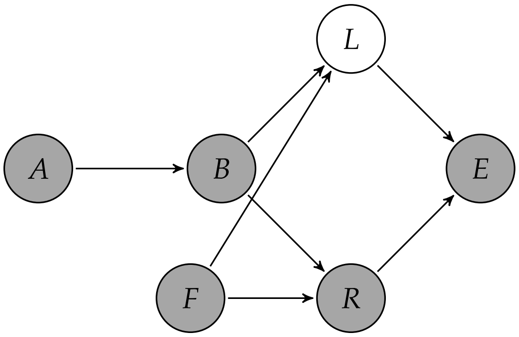


Figure 36: Dependency diagram for basic switch

For Hall (2007, p. 118) writes that the basic switch has 'the obvious causal model' $\langle M, V\rangle$:

$$\begin{array}{|l|}\hline B = A \\ L = \neg F \wedge B \\ R = F \wedge B \\ E = L \vee R \\ \hline A, F, B, \neg L, R, E \\ \hline\end{array}$$

Relative to $\langle M, V \rangle$, F is not a cause of E. For this to be seen, observe that the variables L, R, and E are descendants of F. By condition (3) of the definition of $\gg$, no structural equation except $B = A$ can be removed when looking for an agnostic model with an active path from F to E. Removing it or not, $\langle M', V' \rangle$ is uninformative on F and E only if $B \notin V'$. But then we cannot infer E from $\langle M', V' \rangle [V'][F]$. And so $F \gg E$ fails to hold.

We invite the reader to verify that A is a cause of E, B is a cause of R, R is a cause of E, and B is a cause of E—as desired in the basic switch.[1]

5 Realistic Switch

Some authors are dissatisfied with modelling switches by a *simple switch*, which has been discussed above in this chapter. They believe that any real-world event has at least two causal factors (e.g. Hitchcock (2009, p. 396)). In the train example, for instance, the train can only pass on the right track because everything is normal: nothing blocks the track, it is in good condition, and so on. The simple switch in Figure 30 does not consider such conditions and is therefore considered inappropriate. The following dependency diagram depicts a realistic switch:

[1] Paul and Hall (2013, p. 235) present a modification of their basic switch, where the equation for E is replaced by $E = R$. For the modified scenario, our analysis says that F is a cause of E, as desired. The demonstration is straightforward.

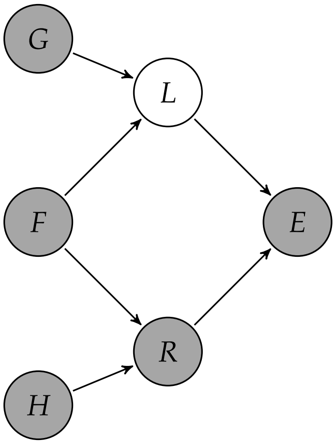


Figure 37: Realistic switch

Flipping the switch (F) and the right tracks being in good working condition (H) bring about the train's travelling on the right tracks (R). And the train's travelling on the right tracks brings about the train's arrival at its destination (E). At the same time, flipping the switch (F) prevents the train from travelling on the left tracks ($\neg L$). However, had the switch not been flipped ($\neg F$), the train would have travelled on the left tracks (L) as the left tracks are also in good working condition (G). And the train's travelling on the left tracks (L) would have brought about the train's arrival (E) all the same. In the actual circumstances, the flipping of the switch determines whether the train travels on the left or right tracks, and so acts like a switch as to the train's arrival. Here is the causal model $\langle M, V \rangle$ of the realistic switch:

$$\begin{array}{|l|}\hline L = G \land \neg F \\ R = F \land H \\ E = L \lor R \\ \hline G, F, H, \neg L, R, E \\ \hline \end{array}$$

There are several causal models $\langle M, V' \rangle$ uninformative on F and E. Some of these have an active path from F to E. Take the model $\langle M, V' \rangle$ for which $V' = \{H\}$. In this model we can infer E from F such that the inferential network is $F \to R \to E$. This is an active path, and so condition (C2) is satisfied.

Our analysis counts the position of a realistic switch as a cause of the train's arrival, which goes against our causal judgements. This observation calls for a refinement of our analysis. In the next section, we explain a potential solution to the problem of realistic switches. This solution, however, reveals another problem: how to delineate between realistic switches and preemption? Our account of deviancy to be developed in Chapter 6 will provide a solution to both problems.

6 Sartorio's Principle

The problem of realistic switches may be solved by adopting the following principle: if an event is a cause of E, then its absence would not be a cause of E in a scenario which otherwise equals the one under consideration. Likewise for absences which are genuine causes. The principle translates to the following refinement of our analysis: C is a cause of E relative to $\langle M, V \rangle$ iff

(C1) $\langle M, V \rangle \models C \wedge E$,

(C2) $\langle M, V \rangle \models C \gg E$, and

(C3) $\langle M, V \rangle \not\models \neg C \gg E$.

Condition (C3) is inspired by Sartorio (2005, p. 90): 'events and their absences would not have caused the same effects'. To be precise, the conjunction of (C2) and (C3) makes the principle explicit that, if an event is a cause of an effect, then its absence would not be a cause of this effect. While the conditional $\neg C \gg E$ may be called a *counterfactual*, it's nonetheless different from the type of counterfactual conditional used in counterfactual approaches to causation.

The above analysis has no problems with realistic switches, given the position of the switch should not count as a cause of the train's arrival. The crucial point is that condition (C3) is violated for the causal claim in question. For the causal model of the realistic switch, explained in the previous section, there is an agnostic model which has an active path from the negation of the candidate cause to the effect in question. Take the model $\langle M, V' \rangle$, where $V' = \{G\}$. In this model we can infer E from $\neg F$ such that the inferential network is $\neg F \to L \to E$. This is an active path. Hence, $\neg F \gg E$, which means that condition (C3) is violated. F does not count as a genuine cause on the present analysis, as it should be.

Unfortunately, condition (C3) leads to a new problem, which concerns the distinction between switches and scenarios of preemption. For this to be seen, let's revisit the simple scenario of early preemption:

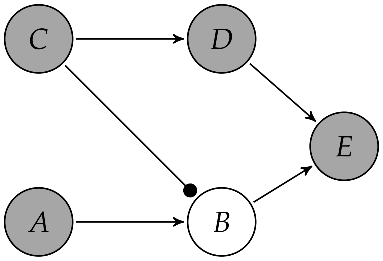
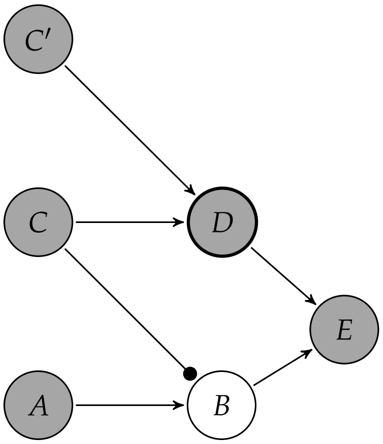


Figure 1: Early preemption

It's easy to show that condition (C3) is satisfied for the causal claim that C is a cause of E. There is simply no agnostic model $\langle M', V' \rangle$ in which E can be inferred from $\neg C$ such that M' contains the equations of all descendants of C. Hence, there is no active path from $\neg C$ to E. However, things are different when we replace the directed edge from C to D by an internal conjunctive scenario in which C and C' are needed to activate D:



$D = C \wedge C'$ $B = A \wedge \neg C$ $E = D \vee B$
$C, C', A, D, \neg B, E$

Figure 38: Extended early preemption

Note that it still makes perfect sense to call C a cause of E. The activation of neuron C contributes to the activation of D, which in turn activates E. The path from A to E, by contrast, is not active since B is blocked by C. We still have a scenario of early preemption.

The present analysis, however, doesn't count C as a cause of E. The problem is that condition (C3) is violated for the claim that C causes E. Take the causal model $\langle M, V' \rangle$, where $V' = \{A\}$. This model is uninformative on C and E. Also, we can infer E from $\neg C$ in this model. The inferential network of the deduction is $\neg C \rightarrow B \rightarrow E$, which is an active path. Hence, $\neg C \gg E$ holds, and so condition (C3) is violated. C doesn't count as a cause of E anymore once we add condition (C3) to our analysis.

Note, furthermore, that the causal model of extended early preemption is structurally indistinguishable from that of the realistic switch. We encounter here an instance of the general problem of isomorphic causal models. On the one hand, two causal models may be isomorphic in the sense of having the same structure. We can obtain the equations of one model from those of the other by a one-to-one mapping between their sets of actual literals and logically equivalent substitutions. On the other hand, our causal judgements are not the same for the two scenarios which the two models were supposed to represent, respectively. In brief, two causal models can be isomorphic, while our causal judgements differ.

Applied to the present case, we can obtain the structural equations of the realistic switch from the equations of extended early preemption by a one-to-one mapping of the actual literals of extended early preemption onto the actual literals of the realistic switch. For example, the equation $R = H \wedge F$ (of the realistic switch) may be obtained from the equation $D = C \wedge C'$ (of extended early preemption) by mapping C onto F, C' onto H, and D onto R. However, we think that C is a genuine case of E in the scenario of extended early preemption, while we do not think that the position of a realistic switch is a cause of the train's arrival. F is not judged to be a cause of E in the realistic switch.

Problems of isomorphism are hard to solve for the following reason. Because of the structural equivalence, there is no way to distinguish between two isomorphic causal models in terms of relations of entailment and deduction. Suppose $\langle M, V \rangle$ and $\langle M^*, V^* \rangle$ are isomorphic. Then it holds that, whenever there is a certain deduction of E from C in $\langle M', V' \rangle$, then there is a corresponding deduction of E^* from C^* in $\langle M^{*\prime}, V^{*\prime}\rangle$, where C^* and E^* have their obvious meanings. Applied to the present case, whenever there is a certain deduction of E from C for the model of extended early preemption, there is a corresponding deduction of E from F for the model of the realistic switch. And vice versa. Hence, there is no way to distinguish between extended early preemption and the realistic switch in terms of inferential pathways without taking further information into account.

We address the general problem of isomorphic causal models in Chapter 6. In line with other approaches to this problem, we take information about normality and deviancy into account in order to better capture our intuitive causal verdicts. This leads to extended causal models $\langle M, V, N\rangle$, where N stands for a set of norms and defaults. Our final analysis is able to solve both problems discussed in this section. First, the problem of realistic switches. Second, the distinction between realistic switches and extended early preemption.

Sartorio's principle is certainly highly plausible and reasonable. But we must acknowledge that a straightforward implementation by condition (C3) has undesirable consequences. For this reason, we have to look for an alternative approach to realistic switches and related problems. Before we come to that, we study a class of causal scenarios for which considerations of deviancy and normality do not matter.

Chapter 5

Entanglement

In this chapter, we show that our analysis captures our judgments in scenarios of entangled causes. Suppose C and A are causal factors of a common effect. We say that C and A are entangled with one another iff we can infer from the presence of one causal factor whether or not the other factor is present as well. A simple case of entanglement is when C is a cause of A, while C and A have a common effect.

We begin with a simple scenario of entanglement in the next section. Then we will give a more principled account of entanglement and the inferential pathways among entangled causes. Thereby, we will give another justification of why we need to retain the structural equations of the descendants of the candidate cause when testing for causation. The remaining sections are dedicated to further causal scenarios of entanglement.

1 Subcause

Recall the scenario of conjunctive causes from Section 2: two causes are necessary for a common effect to occur. We can entangle the two causes by stipulating that one of them is caused by the other. In the resulting scenario, one of the causes necessary for the effect is brought about by the other. Such a scenario is depicted by the following neuron diagram:

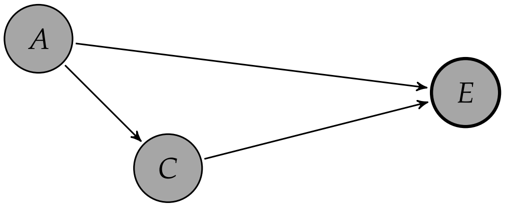


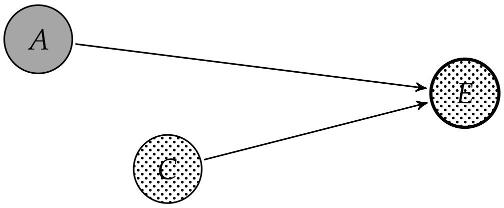
Figure 39: Subcause

Neuron A fires and thereby excites neuron C to fire. Together they bring the stubborn neuron E to fire. Had one of A and C not fired, E would not have been excited. But, crucially, C fires just in case A does. We say that C is a subcause of A. A and C are entangled causes of the common effect E. Our recipe translates the neuron diagram of Figure 39 into the following causal model $\langle M, V \rangle$:

$C = A$ $E = C \wedge A$
A, C, E

Entangled causes are tightly related by structural equations. Here, the subcause C depends directly and exclusively on the *supercause* A. Given the structural equations, the occurrence of E is determined by whether or not A occurs. In this sense, the cause C is subordinate to the cause A.

Relative to $\langle M, V \rangle$, C is a cause of E. There is the following causal model $\langle M', V' \rangle$ which is uninformative on C and E:



$E = C \wedge A$
A

Figure 40: Agnostic model for subcause

We can infer E from $\langle M', V' \rangle[V'][C]$ such that the inferential network of this deduction is $C \rightarrow E$. This is an active path since each inference to a literal depends on the candidate cause C. We represent the path by a thick arrow in the following neuron diagram:

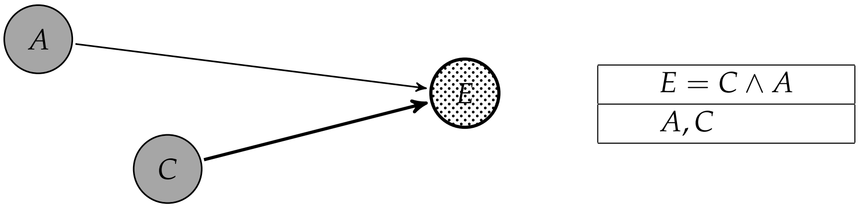


Figure 41: Active path from C to E

We have shown that C is a cause of E. This is the intuitive verdict for two reasons. First, E cannot be caused by A alone. Second, E does not occur without a cause in the scenario. Hence, C should be considered a cause of E. A is also a cause of E, as our analysis says.

2 What is Entanglement?

Before we proceed with further scenarios, let us study the inferential pathways between entangled causes in a more principled way. We started this chapter with a preliminary explanation of entanglement: suppose C and A have a common effect. Then C and A are entangled with one another iff we can infer from the presence of one causal factor whether or not the other factor is present as well. Note that entangled causes may be events or absences. An event may well be entangled with an absence, and vice versa.

The notion of common effect requires some clarification. Suppose C is a direct cause of D, which is a direct cause of E. We could say that C and D have a common effect E, but this understanding is not intended. We avoid the underlying ambiguity by a graph-theoretic explanation of a common effect. Suppose L_C and L_A are two literals, which are true on some causal model. We say that L_C and L_A have a common effect iff there is a variable E such that there is a directed path from C to E and one from A to E, while the two paths have no edge in common.

The presence of a common effect does not imply that the corresponding events or absences are actual causes of this effect. They are mere potential causal factors in virtue of the directed paths to the effect. With these clarifications in place, we can define entanglement.

Definition 5. Entanglement
Suppose $\langle M, V\rangle$ is uninformative on the literals L_A and L_C. Further, suppose L_A and L_C have a common effect. We say that L_A and L_C are entangled with one another iff the value of A determines the value of C, or vice versa, or both. In more formal terms, entanglement means that $\langle M, V \cup \{L_A\}\rangle \models L_C$ or $\langle M, V \cup \{L_C\}\rangle \models L_A$, or both.

Entanglement between two causal factors implies that there is an inferential pathway from one factor to the other. This implication holds because of the completeness of our deductive system and Proposition 1.

We must wonder whether entanglement arises, in general, from the connection via the common effect or some other connection. To answer this question, let us divide the causal model $\langle M, V\rangle$ into two submodels. One is intended to capture the inferential relations of the entangled causes with respect to a common effect. Let us call it $\langle M_E, V_E\rangle$. Another submodel, call it $\langle M_N, V_N\rangle$, captures the inferential relations among the non-descendants of the entangled causes.

To be precise, let M_N be the set of structural equations of those variables which are non-descendants of C or A in the causal graph of M. Note that each variable is a non-descendant of itself. So M_N contains all structural equations of the ancestors of C and A, and, in addition, the structural equations of C and A if there is any. Further, let V_N be the subset of V such that each variable of a literal in V_N occurs in some equation in M_N.

The submodel concerning the inferential relations to a common effect may then simply be defined as the complement of M_N: M_E contains all structural equations of M which are not in M_N. Put differently, M_E contains the structural equations of those variables which are descendants of C or A, and different from C and A. Further, V_E is the subset of V such that each variable of a literal in V_E occurs in some equation in M_E. Thus we have divided a given causal model $\langle M, V\rangle$ into two submodels $\langle M_N, V_N\rangle$ and $\langle M_E, V_E\rangle$. If C and A happen to have more than one common effect, then the set M_E represents the inferential relations with respect to all of these effects.

Which of the two submodels is responsible for the entanglement? It may well surprise us to see that entanglement does not arise from an inferential path via the common effect.

Proposition 6. Suppose L_A and L_C are entangled in the causal model $\langle M, V \rangle$. Let $\langle M_N, V_N \rangle$ and $\langle M_E, V_E \rangle$ be as just explained. Then it holds that

(1) $\langle M_N, V_N \cup \{L_A\} \rangle \vdash L_C$ or $\langle M_N, V_N \cup \{L_C\} \rangle \vdash L_A$, or both, and

(2) $\langle M_E, V_E \cup \{L_A\} \rangle \not\vdash L_C$ and $\langle M_E, V_E \cup \{L_C\} \rangle \not\vdash L_A$.

The first part tells us that entanglement arises from a connection with at least one ancestor of one of the entangled causes. The second part says that the inferential path between entangled causes does not involve any inferences about the common effect. No variable on the inferential path between the entangled causes is a descendant of one of the entangled causes and different from the variables of these causes.

In light of this result, an operation of *disentanglement* falls into place: to disentangle causes which are entangled, if only hypothetically, we need to suspend judgement on some structural equation of the causal model $\langle M_N, V_N \rangle$. This model captures the inferential relations among variables which are non-descendants of at least one of the entangled causes. By contrast, structural equations of the common-effect submodel are not relevant for the entanglement. Hence, there is no need to suspend judgement on these equations. We therefore permit suspension of judgement on structural equations in our analysis, but only with regard to equations of the non-descendants of the candidate cause. This leads to condition (3) of our definition of the epochetic conditional $\gg$: the requirement to retain the structural equations of the descendants of the candidate cause.

In other words, condition (3) leaves room for an operation of disentanglement. This justification complements the justification of this condition in Section 4. We recall the latter justification here briefly. For E to be an effect of C, there must be a model $\langle M, V \rangle$ such that this model is uninformative on E, and E can be inferred from $\langle M, V \rangle [C]$. Suppose there is such a model. Then the variable of E is a descendant of the variable of C. It therefore holds that all potential effects of a candidate cause concern descendants of the latter. By retaining all structural equations of the descendants of the candidate cause, we preserve the inferential relations between the candidate cause and its potential effects. Condition (3) thus ensures that the inferential relations between causes and their potential effects are preserved when we suspend judgement on the candidate cause and its effect.

The operation of disentanglement has been used tacitly in the above section and the previous chapter. Recall the causal scenario of subcauses from the previous section. It is represented by the following causal model: $\langle\{C = A, E = C \wedge A\}, \{A, C, E\}\rangle$. The causal model $\langle\{C = A, E = C \wedge A\}, \emptyset\rangle$ is uninformative on C and A. It is easy to show that C and A are entangled in this causal model. We can infer literal A from literal C, and vice versa. Let us now divide the agnostic causal model $\langle\{C = A, E = C \wedge A\}, \emptyset\rangle$ into two submodels, one of which is about the common effect E, the other is not. The first model is given by $\langle M_E, V_E\rangle = \langle\{E = C \wedge A\}, \emptyset\rangle$. The causal model $\langle M_N, V_N\rangle$ of the non-descendants of C and A is given by $\langle\{C = A\}, \emptyset\rangle$. Clearly, in the context of $\langle\{E = C \wedge A\}, \emptyset\rangle$, there is no way to infer C from A or vice versa. By contrast, in the context of $\langle\{C = A\}, \emptyset\rangle$, we can easily infer literal A from literal C, and vice versa. Both claims of Proposition 6 hold, as desired.

It is furthermore illuminating to look at some causal scenarios of the previous chapter from the perspective of entanglement. Take the short circuit: the boulder's dislodgement (F) threatens to hit the hiker by a rolling boulder (B), and at the same time provokes an action—the ducking (D)—which prevents this threat from being effective ($\neg E$). Recall the causal model: $\langle\{B = F, D = F, E = B \wedge \neg D\}, \{F, B, D\}\rangle$. The following causal model is uninformative on B and D: $\langle\{B = F, D = F, E = B \wedge \neg D\}, \emptyset\rangle$. However, B and D are entangled in this model: we can infer literal B from literal D, and vice versa. Both inferential connections go via the variable F, which is an ancestor of both B and D. The common effect E, by contrast, is not relevant for the entanglement. Again, we can observe that both claims of Proposition 6 hold.

Analogous considerations apply to the switching scenarios in the previous chapter: the variables R and L indicate which route a train is taking on the way to a certain destination. It is easy to show that the literals R and $\neg L$ are entangled in both switching scenarios. They are entangled via inferential pathways which go through the position of a switch, represented by variable F. Notice, finally, that we had to suspend judgement on the structural equation of R in order to show that R is cause of E, the train's arrival at the destination. This suspension of judgement amounts to an operation of disentanglement, as described in this section. Likewise, we had to disentangle the literals B and D in the boulder scenario in order to recognize D (ducking) as a cause of $\neg E$ (the hiker remains unscathed).

We conclude this section with a remark about common causes in the context of Reichenbach's (1956) work on probabilistic causation. Notice that we have a common cause in both the boulder and the switching scenarios. The dislodging of the boulder (F) causes the boulder to roll toward the hiker (B), and it also causes the hiker to duck (D). Likewise, the position of the switch (F) causes the train to go right (R), and it also causes this train not to go left ($\neg L$). With this in mind, we can capture the different behaviour of common causes and common effects by the following observation:

Observation 1. The effects of a common cause are causally connected in the sense that we can infer one from the other. But the causes of a common effect are not connected in this way unless they are connected by a common cause.

This observation may be seen as the deterministic counterpart to an important theorem in Reichenbach (1956, Ch. 19): the effects of a common cause are statistically correlated. But the causes of a common effect are statistically independent of one another—unless these causes have a cause in common.

3 Collaboration

Let us return to the discussion of specific causal scenarios. Beckers (2021, pp. 1361–3) puts forth a series of six scenarios in order to support his causal model account of causation and to challenge others. The latter four scenarios contain entangled causes. The series consists of modifications of a story due to Halpern and Pearl (2005, p. 882). All the scenarios have the following in common. Suppose there is a prisoner and three guards. The prisoner dies just in case guard C loads guard D's gun and D shoots, or if guard A shoots her loaded gun.

In the first scenario, C loads D's gun, D shoots, A shoots her loaded gun, and the prisoner dies. The prisoner's death E is overdetermined by A's shot and the collaboration of C and D. We may as well call it a scenario of *conjunctive overdetermination*. The collaboration scenario itself is not a case of entangled causes, but it gives rise to different scenarios of entanglement to be discussed below.

Note that there is no obvious neuron diagram for this story of collaboration. The reason is that neuron E should be a normal neuron and a stubborn neuron at the same time. E should be normal such that E fires if it gets excited by A. But E should also be stubborn because it should not fire if it receives only one stimulatory signal from either C or D.

Let us therefore represent the causal scenario by a dependency diagram. Recall that such a diagram represents only the dependencies between variables and what events occur. An occurring event is represented by a grey node, an absence by a white node. Unlike a neuron diagram, a dependency diagram does not allow us to read off the structural equations of the causal scenario. Directed arrows, in particular, stand for any type of dependence. A dependency diagram thus contains less information than a neuron diagram.

The dependency diagram of the collaboration scenario shows that the value of the variable E depends on the values of the variables C, D, and A:

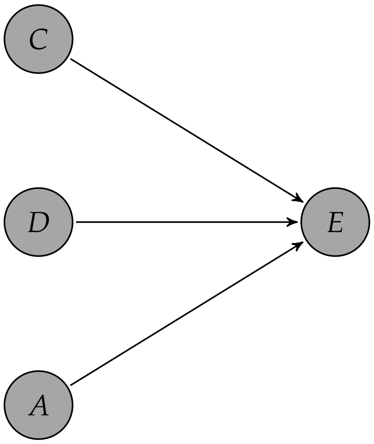


Figure 42: Collaboration

The story can be represented by the following causal model $\langle M, V \rangle$:

$E = (C \wedge D) \vee A$
A, C, D, E

Relative to $\langle M, V\rangle$, C is a cause of E. There is the following causal model $\langle M', V'\rangle$ which is uninformative on C and E:

$E = (C \wedge D) \vee A$
D

We can infer E from $\langle M', V'\rangle[V'][C]$ such that the inferential network of this deduction is $C \rightarrow E$. This is a sequence, and so an active path. Obviously, each inference to a literal depends on C. We have shown that C is a cause of E.

4 Failed Collaboration

In the second scenario, C loads D's gun, but D does not shoot. However, A shoots and so the prisoner dies. Halpern and Pearl (2005) say that C is not a cause of the prisoner's death E in this scenario. Beckers (2021) finds it 'unacceptable' to consider C a cause of E if D does not shoot. And we agree. The story can be represented by the following causal model $\langle M, V\rangle$:

$E = (C \wedge D) \vee A$
$A, C, \neg D, E$

Relative to $\langle M, V\rangle$, C is not a cause of E. We cannot infer E from $\langle M', V'\rangle[V'][C]$ if $\langle M', V'\rangle$ is uninformative on C and E. There are two cases. First, we remove the structural equation of E such that $M' = \emptyset$. However, in the absence of this equation, it becomes impossible to infer E once judgement has been suspended on E. The intervention by the candidate cause C does not enable us to infer E then.

Second, we retain the structural equation of E. Then the following two causal models are uninformative on C and E: $\langle M, \emptyset\rangle$ and $\langle M, \{\neg D\}\rangle$. But we cannot make the inference to E from any of these two models after an intervention by V' and C. The problem is that C is a conjunctive causal factor of E, and so brings about E only together with D. We have thus shown that C is not a cause of E. Again, we have discussed this scenario only in preparation of causal scenarios of actual entanglement to come.

5 Disjunctive Cause Follows Conjunctive Cause

The following scenarios emerge as variations of the scenarios of collaboration and failed collaboration. Beckers sets forth two desiderata for such scenarios. First, C is a cause of E iff D shoots. Second, the relation between D and A should not affect the first desideratum. Our analysis satisfies Beckers's desiderata, as we will show now.

The scenario is like that of collaboration—except that A and D are entangled: A shoots just in case D shoots. The story is modified thus: A observes whether D shoots. If so, A shoots as well; if not, not. Here is its dependency diagram:

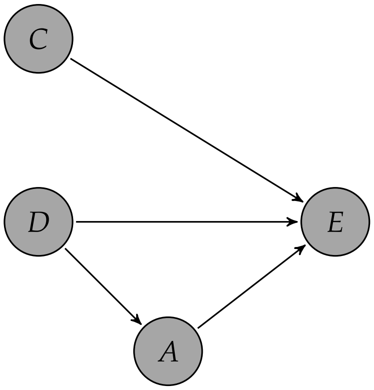


Figure 43: A variant of collaboration

The story can be represented by the following causal model $\langle M, V\rangle$:

$A = D$ $E = (C \wedge D) \vee A$
A, C, D, E

Relative to $\langle M, V\rangle$, C is a cause of E. As for the collaboration scenario, there is the following causal model $\langle M', V'\rangle$ which is uninformative on C and E:

$E = (C \wedge D) \vee A$
D

We can infer E from $\langle M', V' \rangle [V'][C]$ such that the inferential network of this deduction is an active path: $C \rightarrow E$. We have thus shown that C is a cause of E, as desired. By suspending judgement on a structural equation, we have reduced the present causal scenario to one of collaboration. Thereby, we have disentangled the causal factors D and A.

6 Conjunctive Cause Follows Disjunctive Cause

The fourth scenario is like the third scenario—except that the roles of A and D are reversed. The story is modified thus: D observes whether A shoots. If so, D shoots as well; if not, not. Here is its dependency diagram.

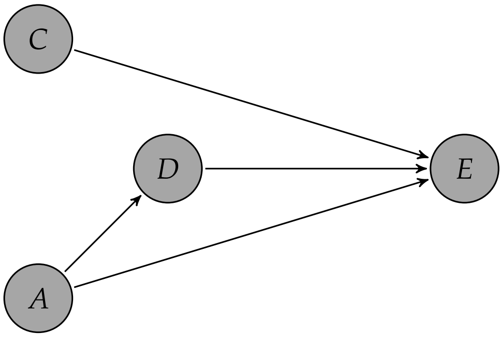


Figure 44: Another variant of collaboration

The story can be represented by the following causal model $\langle M, V \rangle$:

$D = A$ $E = (C \wedge D) \vee A$
A, C, D, E

Relative to $\langle M, V \rangle$, C is a cause of E. Again, the present scenario reduces to that of collaboration. There is the following causal model $\langle M', V' \rangle$ which is uninformative on C and E:

$E = (C \wedge D) \vee A$
D

We can infer E from $\langle M', V' \rangle [V'][C]$ such that the inferential network of this deduction is $C \rightarrow E$. This is an active path since each inference to a literal depends on C. We have shown that C is a cause of E, as desired.

Note that our analysis meets Beckers's desiderata in the scenarios in which D occurs. It says C is a cause of E if D occurs, regardless of the relation between A and D. And it does so by intentionally disregarding the relation between A and D: the structural equation expressing this respective relation is suspended. This suspension allows us to reduce the present causal scenario to that of collaboration.

7 Disjunctive Cause Opposes Conjunctive Cause

The fifth scenario is just like that of failed collaboration—except that A does the opposite of D. The story is modified thus: A observes whether D shoots. If so, A does not shoot; if not, A does shoot. The story can be represented by the following causal model $\langle M, V \rangle$:

$A = \neg D$ $E = (C \wedge D) \vee A$
$A, C, \neg D, E$

Relative to $\langle M, V \rangle$, C is not a cause of E. There is no agnostic model $\langle M', V' \rangle$ for which there is a deduction of E from $\langle M', V' \rangle [V'][C]$ which has an active path. There are three cases. First, we remove the structural equation of E such that $M' = \{A = \neg D\}$. However, in the absence of this equation, it becomes impossible to infer E once judgement has been suspended on E. The intervention by the candidate cause C does not enable us to draw such an inference.

Second, we suspend judgement on the structural equation of A such that $M' = \{E = (C \wedge D) \vee A\}$. Thereby, we obtain the causal model of the scenario of failed collaboration. We have already shown for this model that C is not a cause of E.

Third, we retain all structural equations such that $M' = M$. Then $\langle M, \emptyset \rangle$ is the only causal model which is uninformative on C and E. We can infer E from $\langle M', V' \rangle [V'][C]$ using reasoning by cases with respect to D: if we assume D, we can infer E from C and D. If we assume $\neg D$, we can can infer E via an inference from $\neg D$ to A. The inferential network of this deduction looks as follows:

$$C \to E' \to E \leftarrow E'' \leftarrow A.$$

This inferential network is not an active path: A and E'' do not depend on the candidate cause C. In the same way, we can show that reasoning by cases with respect to A does not give us a deduction of E which has an active path. We have thereby shown that C is not a cause of E, as desired.

8 Conjunctive Cause Opposes Disjunctive Cause

The sixth scenario is just like the fifth—except that the roles of A and D are reversed. The story is modified thus: D observes whether A shoots. If so, D does not shoot; if not, D does shoot. The story can be represented by the following causal model $\langle M, V \rangle$:

$D = \neg A$ $E = (C \wedge D) \vee A$
$A, C, \neg D, E$

Relative to $\langle M, V \rangle$, C is not a cause of E. The argument is similar to the one in the previous section. There is no agnostic model $\langle M', V' \rangle$ such that there is a deduction of E from $\langle M', V' \rangle[V'][C]$ which has an active path. Again, we can can distinguish three cases. First, we suspend the structural equation of E such that $M' = \{A = \neg D\}$. Without this equation, however, it becomes impossible to infer E once judgement has been suspended on E. No intervention with whatever candidate cause—except for E itself—will allow us to infer E from such a model.

Second, we suspend judgement on the structural equation of A such that $M' = \{E = (C \wedge D) \vee A\}$. Thereby, we obtain the causal model of the scenario of failed collaboration. We have already shown for this model that C is not a cause of E.

Third, we retain all structural equations such that $M' = M$. Then $\langle M, \emptyset \rangle$ is the only causal model which is uninformative on C and E. A simple proof by cases enables us to infer E from $\langle M', V' \rangle[V'][C]$: if we assume A, we can infer E directly. If we assume $\neg A$, we can first infer D, and then E from D and C. The inferential network of this deduction looks as follows:

$$D \rightarrow C \rightarrow E' \rightarrow E \leftarrow E''.$$

This inferential network is not an active path: D and E'' do not depend on the candidate cause C. In the same way, we can show that reasoning by cases with respect to D does not give us a deduction of E which has an active path. We have thereby shown that C is not a cause of E, as desired.

The result implies that our analysis meets Beckers's desiderata in the scenarios in which D does not occur. It says C is not a cause of E if D does not occur, regardless of the entanglement between A and D. And it does so by intentionally disregarding the relation between A and D: the structural equation expressing this respective relation is suspended. This suspension reduces the fifth and sixth scenario to the second, namely, the scenario of failed collaboration.

Chapter 6

Deviancy

In this chapter, we address the problem posed by isomorphic causal models. The problem is that there are pairs of scenarios which are structurally indistinguishable for simple causal model accounts, and yet our causal judgments differ (Hall 2007, p. 44). We have already encountered an instance of this problem when comparing a realistic switch with a scenario of extended early preemption in Chapter 4. The problem means trouble for all simple causal model accounts which, like our analysis so far, represent causal scenarios by structural equations and variable values only.

To solve the problem, we amend our analysis. The basic idea is that genuine causes are deviant. Deviancy of an event is explained in terms of violations of a norm or default law. Likewise for deviancy of an absence. We show that the resulting analysis resolves problems arising from scenarios of isomorphic causal models as well as related problems concerning omissions and switches.

1 The Problem of Isomorphic Causal Models

Let us illustrate an instance of the problem of isomorphic causal models. Recall the causal model of the scenario of overdetermination:

$E = C \lor A$
C, A, E

We transform this causal model into an isomorphic one. To this end, negate both sides of the structural equation, which yields $\neg E = \neg C \wedge \neg A$. Then substitute C by F, A by $\neg D$, and E by $\neg E$. This results in the following causal model:

$E = \neg F \wedge D$
$F, \neg D, \neg E$

The two causal models are isomorphic in the sense of being structurally indistinguishable. More technically, we can say that they are isomorphic since the equation of either model can be obtained from that of the other by a one-to-one mapping between their sets of actual literals and logically equivalent substitutions. More generally, we say that a causal model $\langle M, V \rangle$ is isormorphic to another model $\langle M', V' \rangle$ iff M' can be obtained from M by a one-to-one mapping of V onto V' and logically equivalent substitutions.

Notice that $\neg E$ is 'overdetermined' by F and $\neg D$ in the second model, just as E is overdetermined by C and A in the first one. At the level of neuron diagrams, however, we yield a different representation. The second model may be represented by the following diagram:

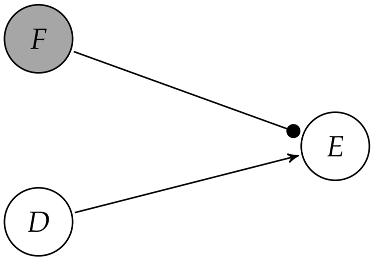


Figure 45: Bogus prevention

Neuron F fires and thereby would inhibit that neuron E gets excited. However, since neuron D is not firing in the first place, there is no danger at all that neuron E gets excited. The prevention of E by F is *bogus*. And so F is arguably not a cause of $\neg E$. The diagram obviously differs from the standard neuron diagram of overdetermination.

Here is a story which fits the structure of bogus prevention. There is an assassin, a potential target, and her bodyguard. The assassin refrains from poisoning target's coffee ($\neg D$), and yet target's bodyguard puts antidote in her coffee (F). Target survives ($\neg E$), of course. Since target's coffee is not poisoned in the first place, there is no danger at all that she dies. The prevention by bodyguard's antidote is *bogus*. And so bodyguard's putting the antidote in her coffee is arguably not a cause of her survival (Hiddleston 2005, Hitchcock 2007).

Overdetermination and bogus prevention are causal scenarios which are structurally indistinguishable for simple causal model accounts. We call a causal model account *simple* if it only factors in structural equations together with values of variables. For simple causal model accounts, there cannot be a structural difference between F in the scenario of bogus prevention and C in the scenario of overdetermination. However, our causal judgments differ. We judge C to be a cause of E in the overdetermination scenario, while we do not judge F to be a cause of $\neg E$ in the bogus prevention scenario.

As the simple causal model accounts of causation, for example Hitchcock's (2001) and Halpern and Pearl's (2005), only factor in structural equations and values of variables, they cannot distinguish between F and C in the isomorphic causal models: C counts as a cause iff F does. This means simple causal model accounts must incorrectly classify the bogus preventer F as a cause in the bogus prevention scenario if they correctly classify the overdeterminer C as a cause in the overdetermination scenario. This is a problem indeed if we take our commonsensical judgments serious.

Our present analysis of causation is a simple causal model account, and so is likewise susceptible to the problem of isomorphic causal models. Hitchcock (2007), Hall (2007), Halpern (2008), Halpern and Hitchcock (2015), and Halpern (2016) all aim to solve the problem by taking into account default and normality considerations. The underlying idea is that the status of genuine causes depends on being deviant from what is normal (Beebee 2004, McGrath 2005). Our solution is guided by the same idea, but comes with a more fine-grained distinction of normality and deviancy. Unlike other accounts using extended causal models, our analysis is able to recognize normal events and norm-compliant actions as causes of other events.

2 Deviancy and Normality

What is it for an event or absence to be deviant? We understand this notion relative to a set of defaults and norms. Roughly, an event is deviant iff it violates a norm or default law. The relevant norms and defaults are rarely made explicit in the informal story of a causal scenario, though.

A default law is one which holds for most of its instances, without however being universally valid. Take the following statement: the tracks of a railway system are in good condition such that trains do not derail. Most people assume this default law to hold when they board a train. They also assume that the locomotive remains operational for the time of their journey. Unfortunately, there are exceptions to these default laws.

Most norms are understood with respect to human behaviour. Keep your promises is a simple example of a moral norm. Don't steel other people's property is both a legal and moral norm. These norms correspond to respective default laws. Most people keep their promises most of the time. Most people respect other people's property. From a logical point of view, it's important that we can express norms and default laws by universal statements.

In engineering, norms are applied to human artefacts. The thread of a bolt needs to satisfy a certain standard. Such norms may be relevant to causal scenarios as well. Roads and railway tracks are required to satisfy certain conditions. Some everyday default laws hold because human artefacts are made and maintained such that certain conditions are met.

Finally, there seems to be an asymmetry between absences and occurring events with respect to their deviancy. By default, an occurring event is more deviant than its absence. Conversely, it's more normal for an event to not occur than to occur. The former rule has been stated by Gallow (2021). Halpern and Hitchcock (2015) assume a similar rule in terms of possible worlds for at least some events. For our approach to normality we assume the following law: by default, an event is absent. We will also refer to this law as the *absence rule*.

At this point, we may already anticipate why a default law about absences could help address certain problems of isomorphic models. Some of these problems are in fact rooted in an asymmetry between absences and occurring events. In a scenario of bogus prevention, the absence of an event behaves logically just like an occurring event in a standard scenario of overdetermination. But only the latter is considered a genuine cause.

In what follows, we will offer two justifications of the default law about absences. One is based on statistical considerations, another draws on the epistemology of absences. Let's begin with statistics.

In general, an event may be described as an object having a specific property at a certain time, where the object is of a certain type. An event of rain may be described as a property of a certain spatiotemporal region. When we say that a neuron is active, we assert that an object, which is a neuron, has the property of being active. Likewise for the event of a kid throwing a rock. Suppose we are asked to consider a certain object of which we only know that it belongs to a certain type. It's a neuron, a human being, a spatiotemporal region, etc. Then we are asked whether the object has a certain property such that the object's having this property amounts to the occurrence of a certain event at a certain time.

Of course, there is almost always no way for us to know the answer to this question. But mere statistical considerations let us infer that the absence of the event in question is more likely than its occurrence. This holds true for virtually all events. Suppose we are told that Amber is a human. Then it's way more likely that Amber doesn't throw a rock at a certain time. Not having any specific information about Amber, it's more likely that she is not a philosopher than that she is. Also, it's way more likely that she does not live in Montreal than that she does. The absence rule holds also for events of rain and the firing of a neuron. Not knowing anything about some arbitrary spatiotemporal region on Earth, it's more likely that it does not rain at a certain time than that it does. Actual Neurons are inactive most of the time. The event of firing is very short as compared to intervals of inactivity.

Of course, there are exceptions to the statistics of absences just asserted. Knowing that Amber is a human, we know that she has a heart. Knowing that an object is a neuron, we know that it is connected with other neurons. For any small spatiotemporal region on Earth, the presence of daylight is at least as likely as its absence. Such exceptions, however, seem to be rare compared to the range of cases for which statistical data favour the absence of an event over its presence. Hence, by default, the default law about absences is well justified for a given type of event. We will adopt this law for all events considered in a causal scenario.

The default law about absences may furthermore be justified by the epistemology of absences. How do we get to believe that an event is absent? It seems fair to say that we notice at least some events relatively directly when they are present. By contrast, we do not notice absences in any direct way. For example, I do not notice directly the absence of an oak tree in my yard. Arguably, my belief that there is no such tree is inferred, implicitly, by the following inference: were there an oak tree in my yard, I would notice it. Since, however, I'm not noticing any oak tree in my yard, there is none. More generally, our means of knowing absences may be reconstructed by the following inference pattern: had event A happened, we would have noticed it. Since we didn't notice A, we think A is absent. The pattern is defeasible since sometimes we overlook events. No doubt, we are not omniscient concerning the occurring events. And it should not be applied to events too small to be noticed in any direct way.

Notably, the default law in question has in fact been used in logic-oriented Artificial Intelligence. It is known there as *negation as failure*. The idea is that, if we cannot infer a sentence A from the respective knowledge base, we take it that A is not true. This is how negation is defined in the programming language *Prolog*. The above inference pattern concerning absences thus becomes part of the meaning of negation. Of course, it is an imperfect way of defining negation since we are not omniscient concerning the occurring events and positive facts. We will not adopt negation as failure as definition of negation here. But we can recognize the underlying inference pattern in everyday reasoning, as shown above with reference to oak trees.

For readers with an interest in the history of philosophy, it may be worth noting that the default law about absences can also be recognized in Leibniz's metaphysics:

> *why is there something rather than nothing?* After all, a nothing is simpler and easier than a something. And moreover, [...] we must be able to give a reason *why they have to exist as they are*, and not otherwise. (Leibniz 1686/1998, Sect. 7, p. 262)

Every existence must have a reason by Leibniz's *Principle of Sufficient Reason*. Non-existence is exempt from this requirement. In the absence of any reason for an existence, we expect non-existence. An existence thus deviates from what we expect by default.

One more clarification concerning default absences is in order. In virtually all causal models we use positive literals for events, while the negative literals stand for absences. But there are exceptions. The position of a switch is a case in point. If we use a propositional variable S for this position, then we should not say that S stands for an event and $\neg S$ for an absence. At least some switches are symmetric. If so, $\neg S$ is not a default.

We are now in a position to define two notions of deviancy of a literal. Let N be a set of norms and defaults.

Definition 6. Weak Deviancy
A literal L is weakly deviant relative to N iff $\neg L$ can be inferred from a consistent subset of N.

Definition 7. Deviancy
A literal L is deviant relative to N iff $\neg L$ can be inferred from a consistent subset of N, but there is no such subset for L itself.

These two notions of deviancy have complementary notions of normality:

Definition 8. Weak Normality
A literal L is weakly normal in N iff L can be inferred from a consistent subset of N.

Definition 9. Normality
A literal L is normal in N iff L can be inferred from a consistent subset of N, but there is no such subset for $\neg L$.

In essence, deviancy is understood as violation of a norm or default. Normality means that a literal conforms to a norm or default. A literal is both weakly normal and weakly deviant iff it violates a norm or default, and conforms to another.

Notice that these concepts of deviancy and normality define a partition on the set of literals. A literal may either be both weakly deviant and weakly normal, or else deviant, or else normal. A deviant literal is weakly deviant but neither normal nor weakly normal. Likewise for normal literals.

Consideration of consistent subsets of N is needed since N may be classically inconsistent. Suppose G stands for the tracks being in good condition.

Then we have both G and $\neg G$ in N. $\neg G$ is in N because of the absence rule. It may seem counterintuitive to maintain that $\neg G$ in N despite the default law that the tracks of a railway network are normally in good condition. However, the absence rule seems deeply entrenched in our causal judgements, as will be shown below.

It's finally worth noting that any occurring event is at least weakly deviant due to the absence rule. No absence is deviant because of this rule. Some absences, however, are weakly deviant. For instance, if the absence of an action violates a promise, then this absence is weakly deviant. Any deviant literal is weakly deviant, but not vice versa. Likewise for normal literals.

The distinction between weak deviancy and deviancy is inspired by the distinction between credulous and sceptical inference relations in formal systems of default reasoning (see, e.g., Meheus et al. (2013)). We will explain below why this distinction is needed for the analysis of some causal scenarios.

Now that we have a basic understanding of deviancy and normality, we can extend our causal models $\langle M, V \rangle$ by an account of deviancy. Let N be a set of default laws and norms, including the default law about absences. This set is supposed to contain the defaults and norms which are presumed to be relevant for the causal scenario in question. We call $\langle M, V, N \rangle$ an *extended causal model*. Such models may be seen as a syntactic counterpart to Halpern and Hitchcock's (2015) extended causal models. The latter work with an order on possible worlds to model normality and deviancy. Such an ordering may be derived from our extended causal models. But we do not need such an ordering to define which literals are deviant.

3 Deviancy of Causes

Let us now exploit the notions of deviancy and normality for our analysis of causation. We set forth two constraints on genuine causes. The first constraint is simple: a genuine cause is always at least weakly deviant. The second constraint is more involved, and concerns the notion of an agnostic model which is supposed to witness an active path from the candidate cause to its effect. Such a model must contain all weakly normal events and absences from the context of the candidate cause in the original causal model. Causal relations are judged against a background of weakly normal events and absences if there are any in the context of the candidate cause. This context is given by the non-descendants of the candidate cause. In more formal terms:

Definition 10. $\langle M, V, N \rangle \models C \gg E$
Let $\langle M, V, N \rangle$ be an extended causal model. $\langle M, V, N \rangle \models C \gg E$ iff there are $V' \subseteq V$ and $M' \subseteq M$ such that

(1) $\langle M', V' \rangle$ is uninformative on C and E.

(2) There is an active path from C to E in $\langle M', V' \rangle[V']$.

(3) All the structural equations of C's descendants are in M'.

(4) C is weakly deviant and any literal $C' \in V \setminus V'$ which is not a descendant of C in M', and different from C, is deviant.

For simplicity, we say that a literal is a descendant of another literal iff the variable of the former is a descendant of the variable of the latter—in the causal graph of the respective causal model. To reiterate, the deviancy condition has two parts. The first part says that the candidate cause is at least weakly deviant. The second part constrains the range of agnostic models which may serve as a witness of an active path. In brief, we cannot suspend judgement on weakly normal events and absences which are non-descendants of the candidate cause.

The first part of our analysis remains unchanged for extended causal models:

Definition 11. Cause
Let $\langle M, V, N \rangle$ be an extended causal model such that $V \models M$. C is a cause of E relative to $\langle M, V, N \rangle$ iff

(C1) $\langle M, V \rangle \models C \wedge E$, and

(C2) $\langle M, V, N \rangle \models C \gg E$.

These two definitions make up our final analysis of causation in Part I. No further additions are needed—as long as we use causal models with structural equations.

4 Bogus Prevention

We have already seen a scenario of bogus prevention in the first section: an assassin does not poison the coffee of a target ($\neg D$). Target's bodyguard administers an antidote (F) and the target survives ($\neg E$). The antidote would prevent target's death in case there had been an attack of poisoning. However, there is no poisoning in the first place which would bring about target's death. Target's survival ($\neg E$) is therefore 'overdetermined' by the lack of poison ($\neg D$) and the presence of antidote (F). For convenience, we copy the neuron diagram for bogus prevention from above:

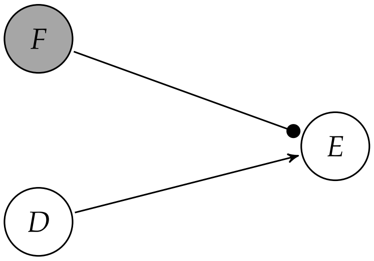
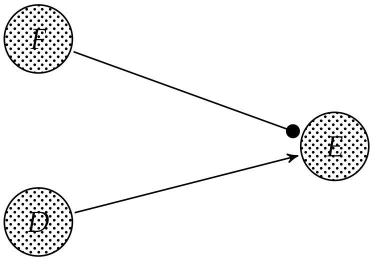


Figure 45: Bogus prevention

Unlike in the classic prevention scenario, there is no threat to be prevented in the first place. This translates to the fact that D is not firing in the neuron diagram. This diagram has an obvious causal model $\langle M, V \rangle$:

$E = \neg F \wedge D$
$F, \neg D, \neg E$

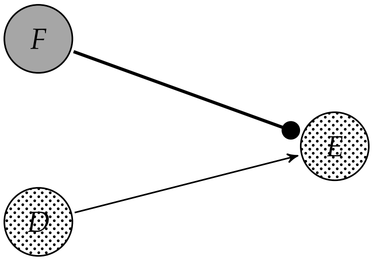
Relative to $\langle M, V, N \rangle$, F is not a cause of $\neg E$ on our analysis. Consider the following causal model $\langle M, V' \rangle$ which is uninformative on F and $\neg E$:



$E = \neg F \wedge D$

Figure 46: Causal model agnostic on F and $\neg E$

This is the only uninformative causal model in the context of which we can infer $\neg E$ from $\langle M, V' \rangle [V'][F]$. For this model, there is a straightforward deduction of $\neg E$ from F whose inferential network is $F \rightarrow \neg E$. This is an active path since each inference to a literal depends on the candidate cause F. Obviously, there is no deduction of the effect $\neg E$ with an active path other than $F \rightarrow \neg E$. The active path $F \rightarrow \neg E$ is visualized by a thick arrow in the following figure:



$E = \neg F \wedge D$
F

Figure 47: Active path from F to $\neg E$

However, the agnostic model $\langle M, \emptyset \rangle$ violates the deviancy condition of the conditional $\gg$ for extended causal models. Note that $\neg D$ is an absence and therefore at least weakly normal. This implies that $\neg D$ is not deviant. Also, $\neg D$ is a non-descendant of the candidate cause F. For these two reasons, we cannot suspend judgement on $\neg D$ without violating the deviancy condition. There is no agnostic model with an active path from $\neg E$ to F other than $\langle M, \emptyset \rangle$. F is therefore no cause of $\neg E$ on our extended analysis, as desired. Administering the antidote does not cause the survival of the target.

5 Extended Bogus Prevention

Consider now the following variant of the above causal scenario. The assassin is a member of a criminal organization, and received an order to poison the target. He fails to carry out the job and is killed (K) because of that by a superior. The scenario may be represented by the following neuron diagram:

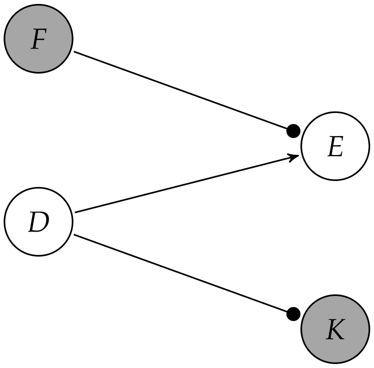


Figure 48: Extended bogus prevention

This diagram has an obvious causal model $\langle M, V\rangle$:

$E = \neg F \land D$ $K = \neg D$
$F, \neg D, \neg E, K$

Arguably, we have still a scenario of bogus prevention here. There is no threat to be prevented by administering an antidote. F should therefore not count as a cause of $\neg E$. Our analysis delivers this verdict indeed. For this to be seen, note that $\neg D$ is an absence, and for this reason weakly normal. $\neg D$ can be derived from a consistent subset of the set N of norms and defaults. However, unlike in the above scenario of bogus prevention, $\neg D$ is not normal since D can be derived from a consistent subset of N as well. After all, the assassin received an order from a superior. Failure to carry out the order violates a norm, even though it may be the right thing to do. Two norms may well be in opposition to one another.

Our solution to bogus prevention remains nonetheless intact. Recall that literals from the context of the candidate cause—-given by its non-descendants in M'—need to be deviant in order for us to suspend judgement on them. $\neg D$ fails to satisfy this condition since it's weakly normal. Hence, the agnostic model $\langle M, \emptyset\rangle$ violates the deviancy condition of the conditional $\gg$ for extended causal models. There is no other agnostic model with an active path from F to $\neg E$. Hence, F is not a cause of $\neg E$ on our extended analysis, as desired.

At the same time, $\neg D$ is a cause of K on our analysis, as it should be. Failure to poison the target causes the assassination of the assassin. There are two agnostic models for which there is an active path from $\neg D$ to K: $\langle M, \{F\}\rangle$ and $\langle M, \emptyset\rangle$. The active path is simply $\neg D \rightarrow K$. Most notably, both agnostic models satisfy the deviancy condition. For this to be seen, note that the candidate cause is not required to be deviant by our analysis. Weak deviancy suffices. $\neg D$ is weakly deviant and weakly normal. Hence, we can suspend judgement on $\neg D$ without violating the deviancy condition. Thus $\neg D$ comes out as a cause of K on our analysis.

Notice that the subtle distinction between weak deviancy and deviancy matters crucially for the present scenario. The deviancy condition of the conditional $\gg$ prohibits suspending judgement on weakly normal events and absences from the context of the candidate cause. Suspension of judgement is only admitted if the event from the context is deviant, and so not even weakly normal. By contrast, the candidate cause itself doesn't have to be deviant. Weak deviancy is sufficient. Hence, we can suspend judgement on $\neg D$ when testing for causation between $\neg D$ and K. But we cannot suspend judgement on $\neg D$ when testing for causation between F and $\neg E$. This is why our solution to bogus prevention remains intact for the present scenario.

In sum, the subtle distinction between weak deviancy and deviancy enables us to capture both of the causal judgments in question: F is not a cause of $\neg E$, while $\neg D$ is cause of K. Administering the antidote is not a cause of target's survival, while the failure to poison the drink is a cause of the assassination of the assassin. Our analysis is the first to achieve this desirable result. Specifically, extant counterfactual approaches fail to capture at least one of the two causal judgements.[1]

6 Bogus Double Prevention

Recall the scenario of double prevention, explained in Section 9 of Chapter 3. There, C is a cause of E because C prevents a threat that would have prevented E. C is a cause in virtue of blocking the threat to E's firing, which

[1] Except for the simple counterfactual account, which, however, is no viable alternative (Andreas and Günther 2025a). The present scenario of extended bogus prevention is inspired by Wysocki (forthcoming).

is initiated by B's firing. Now, let us assume that there is no threat which could be prevented by C. We then obtain a scenario of bogus double prevention (Hall 2007, pp. 119–20):

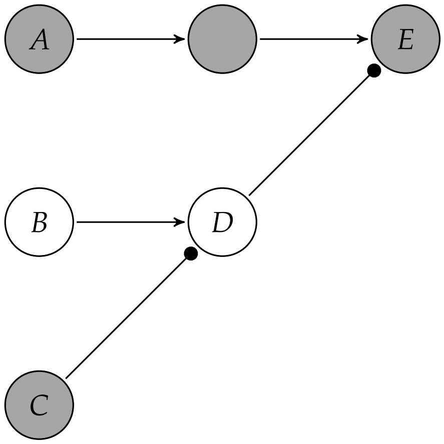


Figure 49: Bogus double prevention

C's firing would have inhibited D's firing if B had fired. However, as B does not fire in the first place, there is no danger that D fires. And so there is no danger that E's firing could be prevented by D's firing. The first prevention is bogus and the second is absent. Hall (2007, pp. 120) thinks it is 'absurd to count C a cause of E', and we agree. Our recipe translates the neuron diagram of Figure 49 into the following causal model $\langle M, V \rangle$:

$D = B \wedge \neg C$ $E = A \wedge \neg D$
$A, \neg B, C, \neg D, E$

Relative to $\langle M, V, N \rangle$, C is not a cause of E. For this to be seen, consider the following causal model $\langle M, V' \rangle$ which is uninformative on C and E:

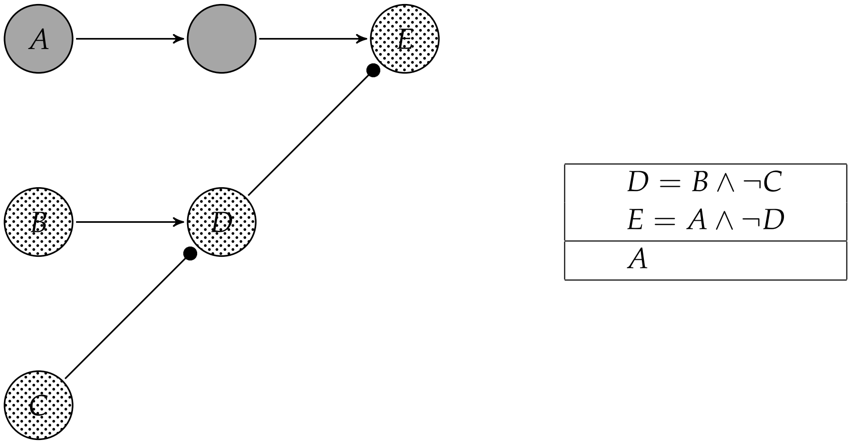


Figure 50: Agnostic model for bogus double prevention

This is the only uninformative causal model in which we can infer E from $\langle M', V' \rangle [V'][C]$. The deduction of E from C may be represented by the inferential network $C \rightarrow \neg D \rightarrow E$. This network is an active path since each inference to a literal depends on the candidate cause C. It may be visualized as follows:

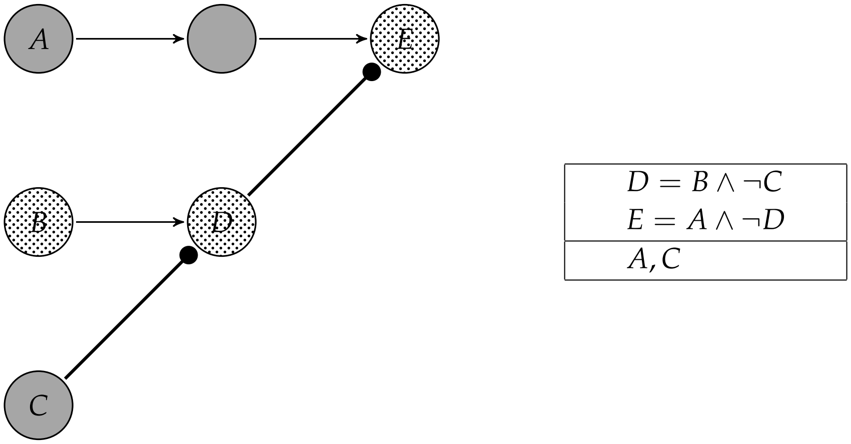


Figure 51: Active path from C to E

However, $\neg B$ is an absence and therefore at least weakly normal. Also, $\neg B$ is a non-descendant of the candidate cause and different from that cause.

For these two reasons, we cannot suspend judgement on $\neg B$ without violating the deviancy condition. There is no other agnostic model in which E can be inferred from C. Hence, C is not a cause of E, as our intuitions have it.

7 Extended Short Circuit

To further our understanding of the deviancy condition concerning the context of the candidate cause, let us consider a variant of the short circuit from Section 1 in Chapter 4. In this scenario, a boulder gets dislodged and rolls toward a hiker. The hiker sees the boulder coming and ducks, so that she does not get hit. If the hiker had not ducked, however, the boulder would have hit her.

Now, we extend this scenario by a variable for a second boulder. No such boulder actually rolls toward the hiker. We want to account for the mere possibility of a second boulder. The second boulder, if dislodged, would come from a different direction so that the hiker would not see it. But the two boulders, if dislodged, would hit the hiker at the same spot. Hence, if the second boulder were dislodged but not the first one, the hiker would not duck and actually get injured. This would be the end of the hike.

The second boulder is an empty threat. Our causal judgements should not change if we take the mere possibility of a second boulder into account.[2] Thanks to the deviancy condition, our refined analysis delivers this result indeed. The structure of the extended boulder scenario can be represented by the following neuron diagram:

[2] See Gallow (2021) for a proposal as to which modifications of a causal model should not alter our causal verdicts. The present scenario is inspired by this proposal.

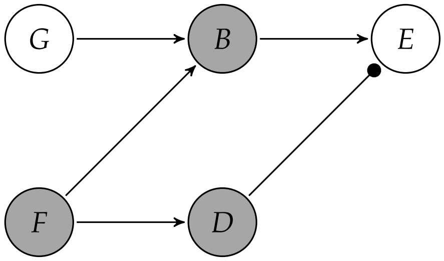


Figure 52: Extended short circuit

This neuron diagram has an obvious causal model $\langle M, V \rangle$:

$B = F \vee G$
$D = F$
$E = B \wedge \neg D$
$F, B, D, \neg G, \neg E$

Unlike the simple short circuit, we now have an agnostic model for which there is an active path from F to $\neg E$. $\langle M, \varnothing \rangle$ is uninformative on both F and E. Also, we can infer $\neg E$ from F in a straightforward manner. The inferential network of the deduction is $F \rightarrow D \rightarrow \neg E$. This is an active path.

However, the agnostic model $\langle M, \varnothing \rangle$ violates the deviancy condition concerning the context of the candidate cause. This model is obtained by suspending judgement on $\neg G$, which is a non-descendant of the candidate cause but not deviant. The absence of a second boulder, represented by $\neg G$, fails to be deviant since N contains the default law about absences. F is therefore no cause of $\neg E$ on our analysis. The empty threat of a second boulder doesn't change the causal verdict about the first boulder, as it should be.

8 Modified Extended Double Prevention

Paul and Hall (2013, p. 198–9) ask us to consider a slight modification of extended double prevention, by adding an idle neuron G:

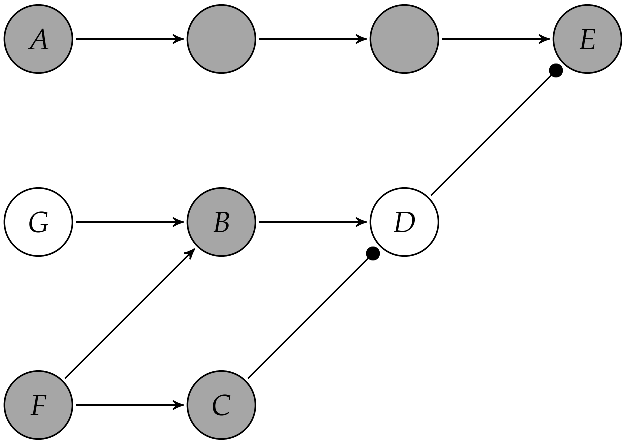


Figure 53: Modified extended double prevention

'Since G does not fire', so argue Paul and Hall, 'the original verdict stands:' F is not a cause of E (p. 199). Our recipe translates the neuron diagram into the following causal model $\langle M, V\rangle$:

$$\begin{array}{|l|}
\hline
B = G \lor F \\
F = C \\
D = B \land \neg C \\
E = A \land \neg D \\
\hline
A, \neg G, F, B, C, \neg D, E \\
\hline
\end{array}$$

There are two causal models $\langle M, V'\rangle$ which are uninformative on F and E such that we can infer E from $\langle M, V'\rangle[V'][F]$, namely for $V' = \{A\}$ and $V' = \{A, B\}$. Indeed, the two agnostic models have the same active path from F to E: $F \rightarrow C \rightarrow \neg D \rightarrow E$.

However, the two agnostic models violate the deviancy condition concerning the context of the candidate cause. $\neg G$ is a non-descendant of the candidate cause F, but fails to be deviant. Therefore, we cannot suspend judgement on this literal without violating the deviancy condition. The reasoning is analogous to the above scenario of an extended short circuit. F is not a cause of E on our analysis, as desired.

9 Isomorphic Modified Extended Double Prevention

Paul and Hall (2013, pp. 198–9) consider a scenario that is isomorphic to modified extended double prevention:

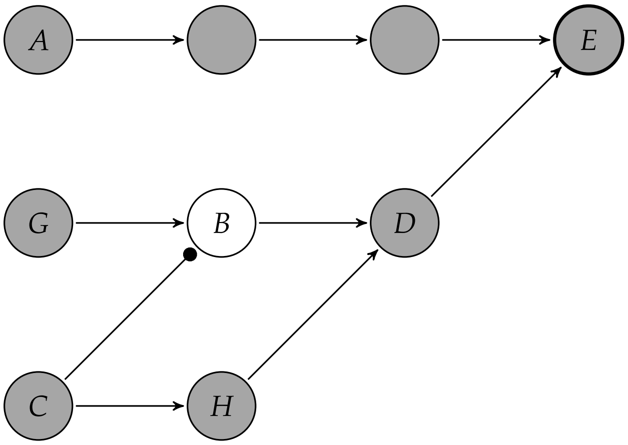


Figure 54: Isomorphic modified extended double prevention

Here, E is a stubborn neuron which requires two stimulations in order to fire. Paul and Hall (2013, p. 199) write 'C clearly is, this time, one of the causes of E—a joint cause with A'. Our recipe translates the neuron diagram into the following causal model $\langle M, V\rangle$:

$B = G \wedge \neg C$ $H = C$ $D = B \vee H$ $E = A \wedge D$
$A, G, C, \neg B, H, D, E$

Relative to $\langle M, V, N\rangle$, C is a cause of E. For this to be seen, consider the following causal model which is uninformative on C and E:

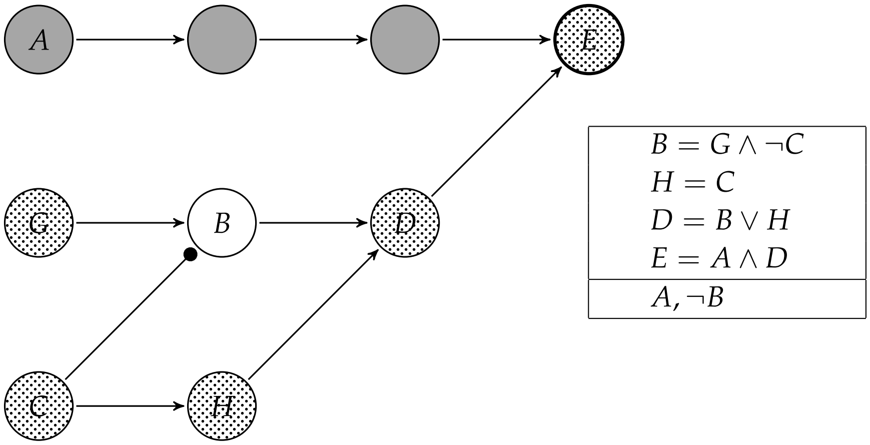


Figure 55: Causal model agnostic on C and E

We can infer E from $\langle M, V' \rangle [V'][C]$ such that the inferential network of this deduction is $C \to H \to D \to E$. This is a sequence, and so an active path. Each inference to a literal depends on the candidate cause C. The active path may be visualized by the following figure:

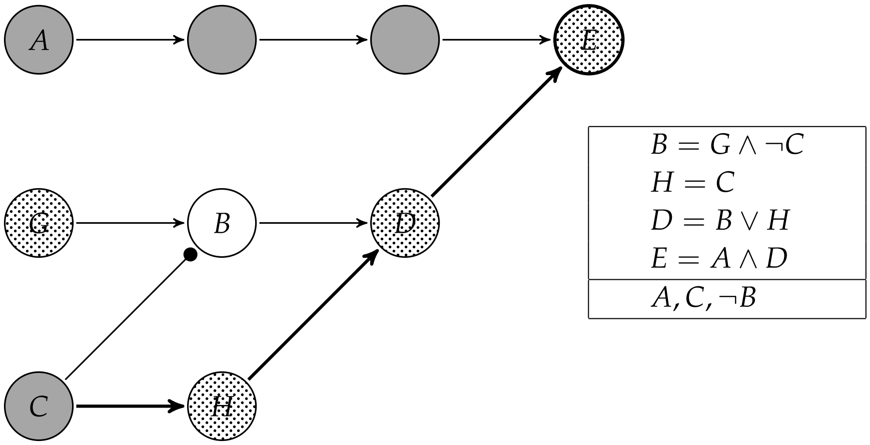


Figure 56: Active path from C to E

This time, the agnostic model satisfies the deviancy condition of the conditional $\gg$ for extended causal models. We obtain this model by suspending judgement on literals which stand for occurring events. There is no norm or default law in place from which we could infer these events. Hence, all literals in $V \setminus V'$, including G, are deviant. Specifically, the literal of the candidate cause and all literals in $V \setminus V'$ which are non-descendants of that cause are deviant. The deviancy condition is therefore satisfied for the agnostic model in question. C is a cause of E on our analysis, as desired.

10 Omissions

Omissions pose yet another problem for many accounts of causation. In a scenario of omission, an event fails to occur and so another event occurs. However, had the event occurred, it would have prevented the other event from occurring (Paul and Hall 2013, p. 174). For example, I go on vacation and ask my neighbour to water my plant. The neighbour promises to water my plant while I am away. However, the neighbour fails to water my plant ($\neg F$), and so my plant dries up and dies ($\neg E$). Had my neighbour watered the plant as promised, the plant would not have died (Beebee 2004, pp. 294–5). The basic structure of omission is given by the following neuron diagram:

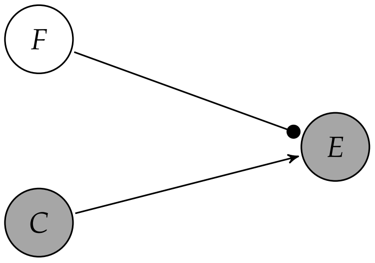


Figure 57: Omission

C fires and thereby excites E to fire. F does not fire. However, had F fired, it would have prevented E's firing. Our recipe translates the neuron diagram of Figure 57 into the following causal model $\langle M, V \rangle$:

$E = \neg F \land C$
$\neg F, C, E$

There is only one uninformative model $\langle M', V' \rangle$ in which we can infer E after an intervention by V' and the candidate cause $\neg F$:

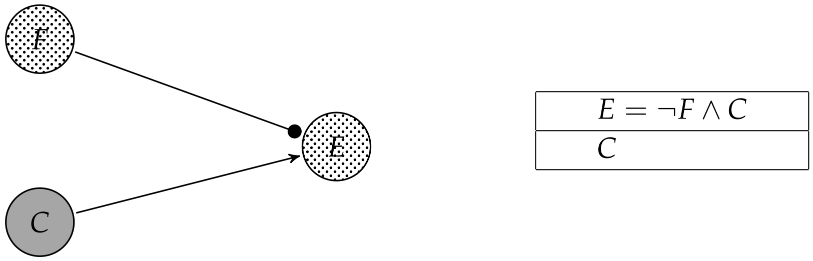


Figure 58: Agnostic model for omission

There is a straightforward deduction of the effect E from the candidate cause $\neg F$ whose inferential network is $\neg F \rightarrow E$. This network is obviously an active path. The deduction may be visualized as follows:

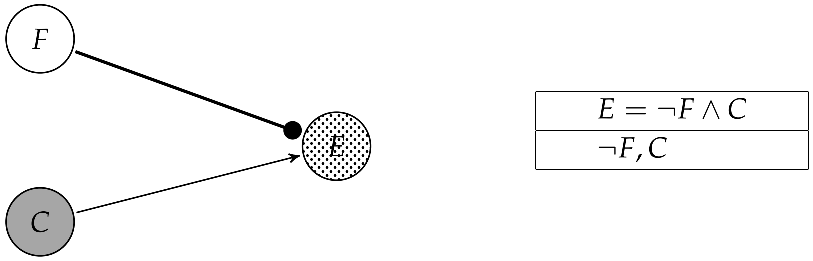


Figure 59: Active path from $\neg F$ to E

In the absence of any norms and promises, $\neg F$ is not a cause of E on our analysis for extended causal models. For $\neg F$ is normal. Hence, the above agnostic model violates the deviancy condition of the conditional $\gg$. For illustration, take the absence that Putin did not water my plant. Since Putin did not make any promises about watering my plant, his failure to do so is rather normal and not deviant at all. His failure to water my plant is not a cause of its death.

The crucial point, however, is that my neighbour promised to water my plant. For this reason, there is a normative expectation that she keeps her promise. In light of this norm, $\neg F$ is weakly deviant. Given that my neighbour promised to do F, the agnostic model in question satisfies the deviancy condition for the candidate cause. The deviancy condition for the context is trivially satisfied. $\neg F$ therefore counts as a cause on our analysis, as desired.

What happens if my neighbour does water my plant? Obviously, the plant doesn't die then. Watering my plant successfully prevents my plant from dying. Notably, our analysis captures this causal judgement as well. Take the causal model $\langle M, \{C\}\rangle$, which is uninformative on F and E. This model has an obvious active path from F to $\neg E$. Also, the agnostic model satisfies the deviancy condition for the candidate cause and its context.

To see why the deviancy condition is satisfied, note that N contains the default $\neg F$ because of the absence rule. Hence, F is weakly deviant, even though it also accords with a norm in N. Both F and $\neg F$ are weakly deviant. The deviancy condition concerning the context of the candidate cause is trivially satisfied. Hence, F comes out as a cause of $\neg E$. In sum, the action of watering my plant and the failure to do so, respectively, qualify as a cause on our analysis when a promise is made. This result accords well with our intuitive causal verdicts.[3]

Extant counterfactual approaches to deviancy struggle to achieve the latter result. The key idea in Halpern and Hitchcock (2015) is that only those counterfactual worlds are considered which are at least as normal as the actual world. If complying with a certain norm is more normal than not complying, this approach is unable to capture the causal efficacy of the norm-compliant action. Such an action cannot be a cause of anything. This result goes against our causal judgements. If a physician, teacher, or police officer is merely doing her duty, the corresponding actions do have causal consequences. A patient gets treated, children learn to read, a suspect is interrogated.

Our syntactic approach to normality and deviancy may well be used to solve the problem for the analysis suggested by Halpern and Hitchcock (2015). We may derive an ordering of possible worlds on the basis of the number of literals—verified by a given possible world—which are normal, weakly normal, weakly deviant, and deviant, respectively. To give an ex-

[3] We are grateful to Christopher Hitchcock for having challenged us about about the general problem of how to recognize the causal dimension of norm-compliant actions in Hitchcock (forthcoming).

ample, a possible world which satisfies strictly more normal literals than another qualifies as more normal than the latter on such an order. The precise details are not straightforward, though, since some literals may not be relevant to the respective causal relation.

Finally, notice that an analogous problem arises for events which are present by default. The presence of oxygen is a case in point. Oxygen is necessary for combustion, but often not mentioned as a cause. Think of a forest fire. It is nonetheless correct to say that the presence of oxygen is a cause of a forest fire, if only a trivial one. Again, the default law about absences is crucial for our analysis to recognize events as causes which are present by default. Thanks to the absence rule, the presence of oxygen is at least weakly deviant. It may therefore qualify as a genuine cause on our analysis.

11 Realistic Switches

We are now in a position to tackle the problem of the realistic switch. Recall the neuron diagram of such a switch from Section 11 in Chapter 4:

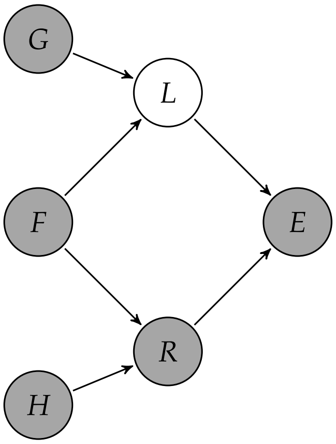


Figure 37: Realistic switch

Further, let us recall that the switch (F) and the right tracks being in good working condition (H) bring about the train's travelling on the right tracks (R). And the train's travelling on the right tracks brings about the train's arrival at its destination (E). This is the causal model $\langle M, V \rangle$ of the realistic switch:

$$\begin{array}{|l|}\hline L = G \land \neg F \\ R = F \land H \\ E = L \lor R \\ \hline G, F, H, \neg L, R, E \\ \hline \end{array}$$

We have seen in Chapter 4 that there are exactly two agnostic models with an active path from F to E: $\langle M, \{H\} \rangle$ and $\langle M, \{H, \neg L\} \rangle$. Both have $F \rightarrow R \rightarrow E$ as an active path. The crucial point, however, is that both models violate the deviancy condition of the conditional $\gg$. By default, we expect the tracks to be in good condition and unblocked. In most countries where trains are taken on a daily basis, railway tracks are in a reasonable condition.

We can therefore assume H and G by default. This implies that H and G are weakly normal, and so not deviant. Further, note that H and G are non-descendants of the candidate cause and different from that cause. Suspending judgement on either of these two literals therefore violates the deviancy condition concerning the context of the candidate cause. This is why the two agnostic models—which have an active path from F to E—violate the deviancy condition. Hence, F is not a cause of E on our analysis, as desired. Flipping the switch does not cause the train to arrive at its destination.

Notice that our analysis has no difficulties recognizing other causal relations of the realistic switch which accord with our causal judgements. F causes R, which means that flipping the switch causes the train to go right. H causes R, that is, the right tracks being in good working condition causes the train to go on these tracks. Notably, H also causes E, which is fairly intuitive. The right track being in good condition causes the train to arrive since the switch directs the train on these tracks. Finally, R causes E, which means that the train going on the right tracks causes the train to arrive.

These results can be obtained in a relatively straightforward manner. Despite being weakly normal, the tracks being in good condition comes out as a genuine cause since causes don't have to be deviant on our analysis. Weak deviancy suffices. Extant counterfactual accounts are in trouble in this regard for reasons explained in the previous section.

Our analysis continues to work for other realistic switches. We will show this for two more scenarios. Hitchcock (2009, pp. 395–6) presents a combination of a realistic and the basic switch from Paul and Hall (2013, p. 232), which leads to the following neuron diagram:

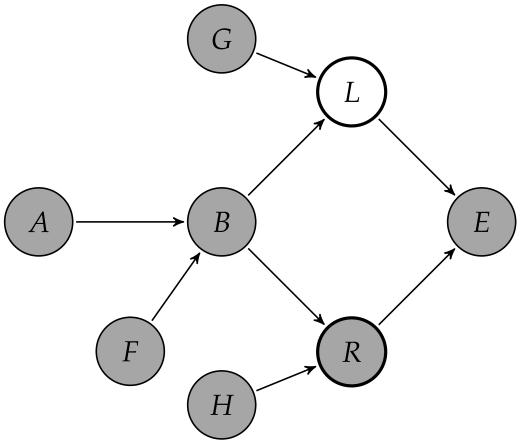


Figure 60: Realistic basic switch

Neuron F acts like a switch as to E. F determines whether the signal coming from neuron A travels to neuron E via neuron R or neuron L. Neurons R and L are stubborn, they will fire only if doubly stimulated. But since neurons G and H both fire, the scenario is very similar to Paul and Hall's basic switch depicted by Figure 35. The realistic basic switch has an obvious causal model $\langle M, V \rangle$:

$$\begin{array}{|l|}
\hline
B = A \\
L = \neg F \wedge G \wedge B \\
R = F \wedge B \wedge H \\
E = L \vee R \\
\hline
A, G, F, B, H, \neg L, R, E \\
\hline
\end{array}$$

The causal verdicts delivered by our analysis for the realistic basic switch are analogous to those of the realistic switch, discussed in the previous section. If G and H are default assumptions, F is not a cause of E. Otherwise, it is. In any case, F and H are causes of R, and R a cause of E, as it should be. The demonstration of these results is analogous to the corresponding demonstration in the previous section.

Whether or not the default assumptions concerning G and H are justified depends on the interpretation of the neurons and the variables in the causal model. Suppose the neurons are taken to represent a scenario of switching with a train and railway tracks in such a manner that G and H stand for the tracks being in good condition and unblocked. Then it is reasonable to assume G and H by default, as has been argued in greater detail in the previous section. Our intuition that switches are not causes of events which occur no matter in what position the switch is derives from scenarios like the train scenario. For such scenarios, it seems justified to assume G and H by default.

12 Looking Back: Preservation of Causal Verdicts

We have shown that our causal model analysis captures our intuitive causal verdicts for virtually all scenarios which received some prominent attention in the literature. All major problems solved, nothing left to do? Not quite. In this chapter, we have refined our simple causal model analysis by a deviancy condition on causes and their context. We must wonder whether this refinement alters any results obtained for the simple causal model analysis, which does not have a deviancy condition. That is, we need to show that our extended causal model analysis—which comes with a deviancy condition—continues to work for the scenarios discussed in the chapters *Classics*, *Non-Transitivity*, and *Entanglement*.

The demonstration is relatively straightforward. Let us begin with causal relations shown to qualify as such on our simple causal model analysis. Suppose $\langle M, V \rangle$ is a causal model. C counts intuitively as a cause of E. Also, our simple analysis recognizes this causal verdict. Further, suppose that the latter result has been established using the agnostic causal model $\langle M', V' \rangle$. Then we know that the causal relation between C and E is recognized by our extended causal model analysis if the following two conditions are met:

(1) All literals in $V \setminus V'$ which are non-descendants of C in M' stand for occurring events.

(2) No literal in $V \setminus V'$—which is a non-descendant of C—can be inferred from a consistent subset of the set N of norms and defaults.

In other words, if an agnostic causal model satisfies these two conditions, then it satisfies the deviancy condition of the conditional $\gg$ for extended causal models. Note that we put the deviancy condition on top of the simple causal model analysis to build the extended one. No other condition is added.

Now, we must show that all the demonstrations that some literal C qualifies as a cause of another literal E—given in the chapters *Classics*, *Non-Transitivity*, and *Entanglement*—do in fact satisfy conditions (1) and (2). Verification of this claim is tedious, but not difficult. It can be simplified by two observations. First, none of the agnostic models $\langle M', V' \rangle$—used in the chapters 3, 4, and 6 to show that some C causes some E—is obtained by suspending judgement on an absence. Second, we did not encounter any causal scenarios with an occurring event D such that D is a non-descendant of the candidate cause and stands for a norm-compliant action. Nor did we discuss a causal scenario in which some event D is a non-descendant of the candidate cause and occurrent by default.

For example, there is no norm or default according to which a kid throws a rock. The neurons of a living organism fire a lot, but not by default at a specific time. The causal scenarios of overdetermination, collaboration, and its variants concern the killing of a prisoner. At least in the absence of further information, such actions are not known to comply with any norm. Quite to the contrary. Hence, the set N does not contain a norm from which the shooting of a prisoner could be inferred. The events in question are therefore deviant.

We still need to check the negative results obtained in the chapters 3, 4, and 6. Suppose we have shown that A does not count as a cause of E on our simple causal model analysis, and this accords with our causal verdicts. The preempted cause in a scenario of preemption is a case in point. Then A does not count as a cause of E on our extended analysis either. Note again that we put the deviancy condition on top of the simple causal model analysis to build the extended one. Hence, if a causal relation does not pass the epochetic test for causation of the simple analysis, this relation cannot pass the test of the extended analysis either.

All problems solved? Unfortunately, we are not done yet. We still need to investigate whether there are any variants of the causal scenarios—studied in the previous chapters—such that the deviancy condition has undesirable consequences.

13 Revisiting Overdetermination and Preemption

Conjunctive scenarios do not pose any problems for our deviancy condition. In such a scenario it suffices to suspend judgement on the candidate cause in order to find an agnostic model with an active path. Things are different for scenarios of overdetermination and preemption. Here we need to suspend judgement on at least two candidate causes. Now, suppose the presence of one overdetermining cause accords with a norm or default, but not the other. Then the former comes out as a genuine cause on our analysis, but not the latter. This result is obtained because weak deviancy suffices for the candidate cause, but not for events in its context in order to suspend judgement on them.

For illustration, recall a standard example of overdetermination, discussed in Chapter 3. A prisoner is shot by two soldiers at the same time, and each of the bullets is fatal without any temporal precedence. Now, suppose one of the soldiers had an order to shoot the prisoner, but not the other. Then the shooting with an order is weakly deviant and weakly normal. The shooting without an order, by contrast, is deviant. If we test for causation of the soldier with an order, all goes well. The candidate cause is weakly deviant and the shooting of the other soldier is deviant. We can suspend judgement on both without violating the deviancy condition. The shooting which accords with an order comes out as a genuine cause, as it should be.

Problems arise when we test for causation of the shooting without an order. Then we cannot suspend judgement on the shooting with an order since the latter is only weakly deviant and in the context of the candidate cause. The shooting without an order does therefore not come out as a cause on our analysis. This is not the desired result. Both soldiers should be held at least causally responsible for their action. The legal aspect is tricky for the soldier who was ordered to shoot.

The problem, however, may be solved by properly extending the causal model used to analyse the scenario. The action of giving an order to one of the soldiers may be made explicit by an additional variable in the model. Then we can suspend judgement on the action of giving an order to shoot the prisoner. This suspension in turn allows us to suspend judgement on the shooting of the soldier with an order—when we test for causation of the shooting without an order. In the absence of explicit information that an order was given, we can assume that no such order was given by the default law about absences. Both events of shooting then count as a cause of the prisoner's death on our analysis. Extending a given causal model is well justified if that model leaves out some information from the informal story of the causal scenario.

However, we must wonder whether the strategy just outlined affects our solution to realistic switches. Arguably, it does not. There is no specific event which causes the tracks of a railway network to be in good working condition. Such tracks need to be maintained in order for them to remain in good condition, but this is done by default. By contrast, there is no default law or norm according to which soldiers in the military are ordered to shoot prisoners of war. Even in a combat situation, there is a specific command to attack the enemy or to fight an attack. We can make that command explicit in the respective causal model.

Notice that the strategy of extending a given causal model—to deal with potential problems of overdetermination—doesn't affect our solution to the problem of omissions. We may well extend the causal model by a variable for my neighbour promising to water my plant. Then we can suspend judgement on that promise. However, we don't have to. One agnostic model with an active path which satisfies the deviancy condition suffices to recognize a causal relation on our analysis. Note, furthermore, that no absence is deviant on our account of deviancy. This property is crucial to our solutions to bogus prevention, extended bogus prevention, bogus double prevention, extended short circuit, and modified extended double prevention. These solutions therefore are not affected by extending the respective causal model as long as the core structure of the scenario remains intact.

Let's move on to causes in scenarios of preemption which are present by default. Suppose Tom wants to become an underwater diver. As a first step, Tom learns how to use an oxygen bottle of the diving equipment outside of the water. When he uses the bottle, it provides his lungs with oxygen. So the bottle may well be considered a cause of his blood getting enough oxygen. The oxygen in the ambient air is only a preempted cause of that effect.

Obviously, we cannot suspend judgement on the presence of oxygen in the ambient air when we test for causation of the oxygen in the bottle. This would violate the deviancy condition concerning the context of the candidate cause. However, it's far from obvious that such a suspension is needed in order to recognize the oxygen in the bottle as a cause. The action of putting on the diving equipment rather acts like a switch with regard to the supply with oxygen. Thereby, Tom switches from the oxygen of the ambient air to oxygen from the bottle. The switch doesn't come out as a cause on our analysis, as it should be. This can be seen from the following variant of the realistic switch:

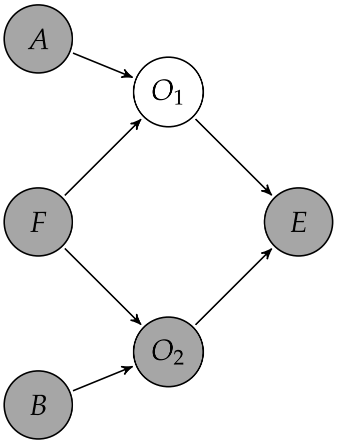


Figure 61: Supply of oxygen

A stands for the presence of oxygen in the ambient air, B for the oxygen in the bottle. O_1 and O_2 stand for the lungs taking in oxygen from the ambient air and the bottle, respectively. The meaning of E is that Tom's blood gets enough oxygen. F means that Tom puts on the diving equipment with the oxygen bottle.

Let's test for causation of F with respect to E. Note that we cannot suspend judgement on A without violating the deviancy condition. Hence, the only agnostic model which is in line with the deviancy condition is $\langle M, \{A\} \rangle$. This model, however, doesn't have an active path from F to E. F therefore doesn't count as a cause on our analysis, as it should be. Putting on the diving equipment is not a cause of why Tom's blood gets enough oxygen.

However, it's easy to show that B is a cause of both O_2 and E on our analysis, which is the desired result. This means that the oxygen bottle is recognized as a cause of Tom's lungs and his blood getting enough oxygen. The causal model $\langle M, \{F\}\rangle$ is uninformative on B and E, has an active path from B to E, and satisfies the deviancy condition. This result holds independently of whether B holds by default since weak deviancy suffices for the candidate cause to satisfy the deviancy condition.

What happens if Tom is not using the diving equipment to get oxygen? Not using this equipment cannot be a cause on our analysis since $\neg F$ is not even weakly deviant. However, A—the oxygen from the ambient air—is then a cause of E. Tom's blood gets enough oxygen because of the ambient air. This is the desired result.

For further illustration, take a scenario about electricity. Suppose a single-family home has solar panels. Starting from a certain threshold, supply of electrical power switches from the grid to the solar panels. The grid provides electricity by default. Not so the solar panels. We can analyse the scenario exactly along the lines of the above scenario about the supply with oxygen from an oxygen bottle. Electricity gained through the solar panels is then recognized as a genuine cause of why the household is supplied with electrical power, given the solar input is above the threshold.[4]

What happens if electrical power comes from both the grid and the solar panels? Arguably, we are then in a scenario of conjunctive causation. At least when we have sufficiently fine-grained information about the amounts of electrical power from the two sources, respectively, our analysis is able to recognize both causal relations. Obviously, non-binary variables are needed then. We have shown in Appendix A how our formalism of causal models may be extended by such variables.

Notably, we have a backup cause in each of the two scenarios just discussed. The oxygen in the ambient air and the power from the grid. For this reason, it makes sense to view them as scenarios of preemption. At the same time, we have shown that an analysis along the lines of a realistic switch makes perfect sense as well. This observation nicely illustrates that we seem to lack a clear delineation between switches and preemption. In the absence of such a delineation, we may have to be prepared that realistic switches overlap with some scenarios of preemption. This is not to say that

[4] The scenario is inspired by Weslake (forthcoming).

the two types of scenarios are the same.[5]

Finally, it remains to consider scenarios of overdetermination and preemption in which the two candidate causes are present by default. We found it very difficult to construct such a scenario. Here is an attempt. Imagine ski resorts were to make artificial snow by default in most regions of the world. Also, they are getting plenty of natural snow by default. Mid winter a ski resort has more than enough snow for skiers to enjoy the runs. This event may be overdetermined by two causes of snow. Our analysis is unable to recognize them as such. That's a clear-cut counterexample. At the same time, it's a silly example whose defaults are contrary to the facts. As a matter of fact, many ski resorts can't rely on natural snow anymore. Those who still can, like the ones in Western Canada not too far from the coast, do not make artificial snow, at least not by default.

In sum, our analysis does face potential problems with scenarios of overdetermination and preemption in which at least one cause is present by default or accords with a norm. For some scenarios of overdetermination we have shown how such problems can be dealt with by an extension of the respective causal model. For two scenarios of preemption we could show that an analysis in terms of a realistic switch is perfectly coherent and so allows us to capture all intuitive causal verdicts. It remains to be seen whether these two strategies work for all scenarios of overdetermination and preemption in which the deviancy condition poses a challenge for our analysis.

One lesson we may have to draw from the challenges observed in this section is that our causal verdicts depend, at least to some extent, on an epistemic perspective. Some causal judgements can be reconstructed from an epistemic perspective which takes certain norms and defaults for granted. For other causal judgements it may be important to disregard certain norms and defaults. We encourage our readers to further challenge and improve our analysis.

[5] Beckers and Vennekens (2018) suggested that the realistic switch qualifies as a scenario of preemption. To our mind, this proposal goes too far, but there seems to be a grain of truth in it. Sartorio (2005) discusses the distinction between switches and preemption at greater length, without however arriving at a general account of switches. So far, our understanding of switches seems to be almost entirely guided by examples which are taken as paradigmatic.

Part II

The Reductive Analysis

Chapter 7

A Humean Analysis

So far, our epochetic analysis of causation relied on causal models with structural equations. Such equations represent certain elementary causal dependences. They are elementary in the sense that no further analysis of the dependence relation is provided. Any attempt at a general explanation of the causal meaning of a structural equation takes some causal or modal notion as antecedently understood, or runs into problems inherent in Lewis's counterfactual analysis of causation.

It's time to dig deeper. It is time to take up the challenge of analysing causation without taking any causal or modal notions as primitively given and antecedently understood. Specifically, we aim at a theory of causation which does not rely on causal laws. In this sense, our analysis of causation in Part II will be reductive. Moreover, we hold on to the logical form of an explicit definition when analysing causation.

How to distinguish between causes and effects in a deterministic setting? To get started, we will reconsider the Humean convention: a cause precedes its effect in time. In essence, we define that C is a cause of E—relative to an epistemic state—iff the following three conditions are met. First, C and E are occurring events. Second, after suspending judgement about C and E, we can infer E from C. And third, C precedes E.

This analysis is still preliminary since it does not address well-known problems of the Humean approach to the direction of causation: spurious causation in common-cause scenarios, simultaneous causation, and backward causation. This chapter begins with an overview of how we will solve and address these problems. Then we will study key concepts of belief revision theory in order to define an epochetic conditional for epistemic states with beliefs about laws and presumed facts. The laws and presumed facts come with information about temporal relations, yet are free of any causal notions. We will begin with a minimalist and syntactic notion of law such that our theory does not rely on any substantial notion of law of nature.

In the subsequent chapters, we will then refine our Humean analysis such that the problems of spurious and simultaneous causation are taken care of. In the final chapter of this part, we will accommodate the conceptual possibility of backward causation. We achieve this by a disjunctive approach to the direction of causation along the lines of extant accounts of backward causation.

1 The Humean Convention

We set out to analyse causation, starting from what we called an *epochetic conditional*:

> $A \gg B$ if and only if, after suspending judgment about A and B, we can infer B from the supposition of A.

We take this conditional to help us understand what it is for a proposition to be a reason for another proposition. The notion of reason is thus understood in a broadly logical way: reasons are inferential in nature. Given A and B are believed, A is a reason for B—relative to our respective epistemic state—iff B is inferable from A in the context of a set of other propositions which we continue to believe after judgement has been suspended about A and B. Propositions are assumed to have a sentential representation.

This analysis of the notion of reason has still some shortcomings, which may be referred to as *symmetry problems*. Sometimes A comes out a reason for B and so does B for A, while intuitively we consider only one of the two propositions as reason for the other. Fortunately, such symmetry problems disappear in our reductive analysis of causation, as will become obvious below.

Reasons are different from causes. At least some reasons are not causes of what they are a reason for. A case in point are reasons for beliefs in mathematics. To analyse causation using some notion of reason, we therefore need to explain how we can distinguish between reasons which correspond to causes and those which do not.

How do we get from reasons to causes? Our strategy throughout this investigation is to impose further constraints on the inferential relations between antecedent and consequent of the conditional $\gg$. In Part I we have strengthened the definition of $\gg$ by the framework of causal models and an inference relation in such models. Together with a few more refinements, this additional constraint led to a relatively simple and surprisingly powerful analysis of actual causation.

How do we get from reasons to causes without structural equations? We begin our journey by reconsidering the Humean convention: a cause precedes its effect in time. We are well aware that this convention has come out of fashion in philosophy.[1] Our reconsideration therefore needs to be well justified. Let's begin with a review of the main reasons why the convention has been rejected:

(1) The Humean convention is violated in some common-cause scenarios of spurious causation.

(2) It is also violated in causal scenarios of simultaneous causation.

(3) It rules out a priori the conceptual possibility of backward causation.

(4) The convention makes the direction of causation a matter of definition.

(5) It blocks any account of the direction of time in terms of the direction of causation.

All of these arguments will be carefully considered in our analysis. The three subsequent chapters are dedicated to problems (1) to (3), respectively. For now, let us briefly indicate how our final analysis addresses these problems.

[1] The ranking-theoretic analysis by Spohn (2006, 2012) seems to be the only prominent exception to this trend.

As in Part I, we will impose further constraints on the inferential relation between antecedent and consequent of the conditional $\gg$. The first constraint is a proof-theoretic variant of the Humean convention: any inferential step—on the inferential path from the antecedent to the consequent—must be forward-directed in time if the inferred sentence asserts the occurrence or absence of an event. Second, all laws used on the inferential path from the antecedent to the consequent must be non-redundant. These two constraints will eventually allow us to tell spurious and genuine causes apart from one another.

As regards presumed scenarios of simultaneous causation, the Humean convention helps us recognize an explanatory asymmetry between cause and effect. In essence, the cause in a relation of simultaneous causation has a Humean causal explanation which is independent of the effect, but not vice versa. This observation allows us to account for scenarios of simultaneous causation within a broadly Humean approach.

The notion of backward causation has remained controversial in the literature. Despite substantial efforts, we are far away from any consensus on how this notion could be understood properly. Backward causation refers to causation which violates the Humean convention: causation where the effect precedes its cause in time. It remains unclear what empirical findings could count as empirical evidence for instances of backward causation in our world. It remains likewise unclear what it means that some event causes another if the Humean convention is dropped altogether.

We will review prominent counterfactual alternatives to the Humean convention in Chapter 10. Drawing on Reichenbach (1956), we will then show that Woodward's interventionist account implicitly relies on the Humean convention. Lewis's attempt to derive the direction of causation from the semantics of counterfactuals will be shown to run into seemingly insurmountable problems. Our assessment and criticism complements extant critical work on Lewis's attempt.

Should backward causation then be rejected outright because causation is so difficult to understand without the Humean convention? Not quite. Despite appearances to the contrary, we do not have to choose between the Humean convention and the conceptual possibility of backward causation. There is middle ground between the two. Dowe (2000, Ch. 8) has developed a disjunctive approach to the direction of causation in the framework of Reichenbach (1956). While Dowe's proposal is not intended to combine backward causation with the Humean convention, it may well be used to do so. Consequently, we will outline how backward causation may be understood within our broadly Humean approach. The key idea is that the Humean convention is but one of two means to distinguish between causes and effects.

It is furthermore striking that Price's (1996) account of backward causation is based on a disjunctive approach to the direction of causation as well. The Humean convention remains in place in a restricted and modified form. Our study of the literature on backward causation thus yields a surprising result: prominent, viable accounts of causation which are aimed at capturing both forward and backward causation retain the Humean convention in one form or other. This is an important motivation for our reconsideration of the convention.

To address objection (4), we suggest a Poincarean view of the Humean convention. Maintaining that causes precede their effects is neither an arbitrary stipulation nor a purely factual statement. It's rather comparable to a Carnapian postulate which determines, in part, the meaning of both causation and temporal order. Nonetheless there remains a conceptual order between the two concepts in the sense that causation is more theoretical than temporal order. Details will be worked out in the next section. Objection (5) is motivated by Kant's proposal to reverse the presumed conceptual order of the Humean convention. We will likewise discuss it in the next section.

Our main goal for Part II is to develop a unified and reductive theory of deterministic causation which is extensionally adequate with respect to our causal judgements in science and everyday life. Such a theory may well be developed without answering the question of why the Humean convention is so important for our notion of causation.[2]

[2] See Price (forthcoming), for a recent proposal to answer this question from an interventionist perspective. We see some interesting connections between epochetic conditionals and Price's interventionist approach. Price (1996) distinguishes between hypothetical, counterfactual interventions concerning the past and potential interventions concerning the future. The epistemic situation for the latter is such that we do not have knowledge yet as to whether the cause and the effect occur.

2 Poincarean Conventions

Does the Humean convention make the direction of causation a matter of definition? Well, it depends on how we understand the notion of convention. Commonly we think that conventions could be chosen differently without severe consequences. Using a specific convention is a relatively arbitrary choice. People in the United Kingdom could have chosen to drive on the right-hand side of the road without severe difficulties, at least at a certain time in history. However, there is a more subtle and richer notion of convention to be found in Poincaré's philosophy of science.

Poincaré (1902/1952) realized that certain laws of physics as well as certain assumptions concerning measurement of space and time escape any direct verification and justification. Moreover, Poincaré tried to show that established formulations of laws of physics as well as assumptions about spacetime are guided by some principle of simplicity. Different formulations and assumptions are conceivable, which notably would capture the same observable facts. Poincaré was not in principle opposed to using non-Euclidean geometries in physics. He just thought that any non-Euclidian formulation of mechanics would be more complicated than its Euclidean counterpart. This raises the following question: are the laws of physics and assumptions concerning spacetime mere arbitrary stipulations or factual statements? Poincaré's answer was neither.

Looking from the perspective of Quine (1961), we can say that Poincaré already realized that the laws of physics are neither analytic nor synthetic. While a Poincarean convention has an analytic and a synthetic dimension, it is neither purely factual nor arbitrary. It is a non-arbitrary convention which has factual content in the context of other conventions. It thus very much resembles a postulate in the sense of Carnap's (1958) semantics of theoretical terms. Such postulates have a twofold function: first, setting forth the empirical content of a scientific theory. Second, determining the meaning of theoretical terms. Notably, Poincaré (1902/1952, p. 90) himself literally claimed that certain scientific propositions acquire meaning only by virtue of adopting certain conventions.[3]

What is the benefit of a more nuanced understanding of the notion of con-

[3] For a detailed study of substantial commonalities among the semantic views of Poincaré (1902/1952), Quine (1961), and Carnap (1958), see Andreas (2010).

vention for our analysis? The standard reading of a Humean approach to causation assumes that temporal relations are epistemically accessible without causation. We don't need a Humean theory of causation to make judgements about temporal relations. Put differently, there is a clear-cut conceptual order between time and causation: time is less mysterious than causation so that it can be used as basis for an analysis of the latter. To be precise, the concept of temporal succession is less mysterious than the direction of causation since we can determine temporal relations among events without information on their causal order.

However, the mere dictum that causes precede their effects does not imply any conceptual order between causation and time. We might as well view time and causation as theoretical concepts whose meanings are intertwined and in part determined by a Humean theory of causation. Most of the time, our causal judgements seem to be based on a prior determination of temporal relations. But there may be exceptions. Special relativity seems to be a case in point: based on the postulates that causal interactions need to be mediated by a signal and that no causal signal can travel faster than light, we come to discriminate between the past and future of an event, and an area of spacetime which is neither in the past nor in the future of the respective event.[4]

Our semantic views about causation are inspired by Ramsey (1931b), Carnap (1958), and Sneed (1979), whose overall research programme concerning theoretical terms may be described as follows: we can analyse our understanding of the theoretical terms in a theory T in terms of axioms and what we can do with the axioms in the context of information which is not theoretical relative to T. A piece of propositional information is considered not theoretical relative to a theory T iff this theory is not used to obtain that information, not even implicitly by the use of certain measurement methods.

With this understanding of theoreticity in mind, we aim to analyse our causal judgements in terms of axioms and inferences which use information which is not theoretical relative to our theory of causation in at least a large subset of applications of this theory. No single causal or modal judgement needs to be taken as primitively given. No direct perception of causes

[4] This delineation of past and future is expressed by Minkowski spacetime. The details are complex, which is why we do not attempt to give an introduction to relativistic physics here.

needs to be assumed. In a similar vein, Schurz and Gebharter (2016) have shown how probabilistic causation may be understood as a theoretical concept in the framework of Sneed (1979). The Humean convention is an optional axiom in their reconstruction.

If we determine temporal relations on the basis of causal relations, the latter are based on causal hypotheses which are inferred from Humean causal relations. Such hypotheses are inferred from causal relations which are recognized as such using the Humean convention on the standard reading. Arguably, special relativity is a case in point. There, we take it for granted that propagation of light is caused by emission from sources rather than by absorption. In technical terms, 'waves and radiation fields exhibit a temporal asymmetry in the presence of wave sources' (Frisch 2014, p. 167).[5] Special relativity does not imply a violation of the Humean convention. It only suggests a partial reversal of the conceptual order standardly assumed for this convention.

If we were to accept a complete breakdown of the presumed conceptual order between time and causation, could we still pursue our investigation? This may well be possible. But our analysis of causation would then lose much of its epistemological motivation. We use the term *analysis* in the spirit of Russell (1918/2010) and Carnap's related concept of explication in his (1950). Much of Russell's and Carnap's philosophical work is driven by the methodological principle that there is some level of knowledge and information which is less theoretical and less mysterious than the level to be analysed and explicated. Some residual empiricism—however refined, relativized, and modified by holistic elements—is needed to read our theory as an epistemological analysis of causation.

References to theoretical terms may thus remind us that we have nowadays logical means of analysis which go well beyond a simple reductionism, without however giving up the empiricist spirit in the early work of Russell and Carnap altogether. If we accept a partial reversal of the Humean conceptual order between time and causation, this should make us question a simple adoption of a Humean analysis of causation. We should then not assume that temporal relations among events are completely determined and given without causal knowledge and causal hypotheses. There re-

[5] This phenomenon is also called *asymmetry of radiation*. See Frisch (2014, Chs. 5 and 7) for a detailed discussion.

mains nonetheless a wider basis of the analysis where temporal relations are accessible without any causal knowledge.

In sum, a Poincarean view of the Humean convention has merit for the following two reasons. First, we can use this convention in a theory of causation without assuming any conceptual order between time and causation. Temporal and causal order are two theoretical concepts whose meanings are intertwined completely. Second, we can admit a partial reversal of the conceptual order between the directions of time and causation, which implies a more moderate deviation from the standard reading of the Humean convention. In our theory of causation we pursue the second option. For lack of space, we do not further investigate which role the Humean convention plays in fundamental physics.

Our analysis of causation in Part II qualifies as reductive in two dimensions. First, no causal or modal notions are taken as primitive and antecedently understood. Second, we hold on to the logical form of an explicit definition when analysing causation. Our final analysis—which merges the analyses from Part I and II—remains reductive in the sense that no causal or modal notations are taken as antecedently understood. This analysis has the form of an explicit definition too.

There remains to discuss the reversal of the standard conceptual order between time and causation. In his *Critique of Pure Reason*, Kant (1781/1998) argued that recognition of temporal relations presupposes judgements about causal relations.[6] Knowledge of causal relations is needed to determine temporal relations. This way, he tried to refute Hume's scepticism about the lawfulness of nature. As influential as Kant's transcendental philosophy has been, his causal theory of temporal order has not been very well received. More importantly, a theory of temporal order may well be based on causation by means of the Humean convention. While such a theory reverses the presumed conceptual order of the Humean convention completely, it is still consistent with this convention. Kant did not so much question the Humean convention, but turned its presumed conceptual order on its head.

To our mind, Kant did not get the epistemology of temporal judgements right. Put more carefully, highly sophisticated interpretations of Kant's

[6] See *Analogies of Experience* in Section III, Chapter II, Book II of Division I in Kant (1781/1998).

transcendental analytics are needed to square his views on causation with our judgements concerning temporal order. For it seems as if we can see bodies falling to the ground—rather than rising to the sky—without any law of Aristotle's, Galilei's, or Newton's physics. If such laws were implicitly used when determining temporal order, what are our reasons for not using the opposite laws, which say that bodies rise to the sky? Causal knowledge may well sometimes be used to determine temporal relations, but a complete reversal of the Humean conceptual order is difficult to understand.

3 Belief Revision Theory

Belief suspension is at the centre of our epochetic approach to causation. The theory of belief revision—which can be used to define belief suspension—has been in the background of our approach from the outset. It is time to take a closer look at this theory. We will now explain key concepts of belief revision theory in a more explicit manner. This will enable us to define an epochetic conditional in a manner which is more general than the definitions of $\gg$ in Part I. The latter are confined to causal models $\langle M, V \rangle$. We now want to explain the conceptual and logical foundations of epochetic conditionals in general.

The framework of AGM-style belief revision theory is syntactic: beliefs are represented by sentences in the sense of a formal logical language. The overall objective is to study how sets of beliefs change when some new piece of information is received. Such a study aims to capture both normative and descriptive elements of belief change. AGM is an acronym which stands for the authors of the landmark paper 'On the Logic of Theory Change' by Carlos Alchourrón, David Makinson, and Peter Gärdenfors (1985).

In the wake of the original AGM theory, semantic approaches to belief revision have been developed by Grove (1988), Spohn (1988), van Ditmarsch et al. (2008), and many others. The key idea is to represent epistemic states by sets of possible worlds and a plausibility ordering on such sets. Semantic and syntactic approaches to belief revision have their genuine benefits, respectively. We use a syntactic approach in terms of belief bases—to be introduced below—for mainly two reasons. First, it allows for a concise representation of concrete causal scenarios. Second, belief bases help us solve problems of spurious causation in a manner which is not available to semantic approaches. Of course, we assume that the sentences in a belief base have some intuitive meaning, which however is not further specified. Sometimes we speak of propositions when referring to intuitively meaningful sentences, following the terminology in Gärdenfors (1988).

Suppose K is a set of sentences which represent the beliefs of an agent, while A is a sentence which represents a single belief. In the AGM framework, we have three types of belief change concerning a belief set K by a sentence A:

(1) Expansion $K + A$

(2) Revision $K * A$

(3) Contraction $K \div A$.

An expansion of K by A consists in the addition of a new belief A to the belief set K. This operation is not constrained by any considerations as to whether the new epistemic input A is consistent with the set K of present beliefs. Hence, none of the current beliefs is retracted by an expansion. $K + A$ designates the expanded belief set.

A revision of K by A, by contrast, can be described as the consistent integration of a new epistemic input A into a belief system K. If A is consistent with K, it holds that $K * A = K + A$, that is, the revision by A is equivalent to the expansion by A. If, however, A is not consistent with K, some of the present beliefs are to be retracted as a consequence of adopting the new epistemic input. $K * A$ designates the revised system of beliefs.

A contraction of K by A, finally, consists in retracting a certain sentence A from the presently accepted system of beliefs. This operation will be used to define the suspension of judgement about A for our epochetic conditional. It is crucial that contractions are guided by some principle of minimal mutilation: when retracting a belief A from K, we should hold on to as many of the beliefs in K as possible, without however retaining a subset of K which implies A. $K \div A$ designates the belief set after the retraction of A.

Belief changes can be defined in various ways. A large number of different belief revision schemes have been developed in the spirit of the original

AGM theory. We will assume that epistemic states are represented by *belief bases*. A belief base H is a set of sentences which represent the explicit beliefs of an agent. Belief base revision schemes are guided by the idea that the inferential closure of a belief base H gives us the belief set $K(H)$ of H:

$$K(H) = Inf(H).$$

$K(H)$ contains all beliefs of the epistemic state H, that is, the explicit beliefs and those beliefs which the agent is committed to accept because they are inferable from the explicit beliefs. Inf is an inferential closure operation which contains classical logic. We assume that Inf is given by the consequence operation Cn of classical logic. Hence, the belief set $K(H)$ is defined by $Cn(H)$. By definition, a belief set is logically closed, while a belief base is normally not. To be precise, a belief base need not be logically closed.

For our preferred approach to belief revision, it is crucial that revisions may be defined in terms of contractions and expansions:

$$K * A = (K \div \neg A) + A. \qquad \text{(Levi identity)}$$

Once we have retracted $\neg A$, we obtain a belief set K' which is consistent with A. Hence, we have $K' * A = K' + A$. The Levi identity has been used to define the revision of both belief sets and belief bases.[7]

It is sometimes helpful to distinguish between the belief system K and the epistemic state S which underlies it. Henceforth, we shall make this distinction, and write $K(S)$ for the belief set K of the epistemic state S. Epistemic states contain further information about the beliefs of an agent. In particular, they include some ordering of epistemic priority on such beliefs. The rationale for such an ordering is as follows. Some beliefs are more firmly established and more deeply entrenched than others. A case in point are the laws of scientific theories which have been well confirmed and which are being applied widely and successfully. If we are forced to retract some of our beliefs—because the new epistemic input is not consistent with the total body of our currently held beliefs—we should hold on to our more firmly established beliefs as much as possible, and try to regain consistency by giving up less entrenched ones.

In our theory of causation, we work with two types of belief bases. One has just two levels: the upper level, containing the laws which describe

[7] For a comprehensive introduction to belief revision theory, the reader is referred to Gärdenfors (1988) and Hansson (1999). Hansson (1993) originated the study of belief bases.

relations among events. And the lower level, which contains beliefs about presumed atomic facts. We have used another type of belief base in our account of deviancy: it contains—in addition to laws and facts—certain default laws and default assumptions. The latter type of belief base is not needed for the reductive analysis in Part I, though. Here is a visualization of the simpler type of belief base:

Laws
Facts

Table 1: Belief base with two levels

The levels of epistemic priority affect the determination of belief changes: when we retract a belief A, we first retract atomic beliefs which imply A in the context of the laws. If necessary, we also retract one or several laws, but only if the retraction of A cannot be achieved by retractions of beliefs with lower epistemic priority.

Note that our causal models $\langle M, V \rangle$ may be read as belief bases with exactly two levels. M is a set laws, which have the logical form of structural equations. V is a set of presumed facts. Belief revision theory was already in the background in Part I. We left the belief revision foundations implicit in this part to make our causal model analysis of causation more accessible.

We understand the notion of law in a minimalist and syntactic way: any universal sentence and any implication explicitly believed to be true—by the respective epistemic agent—is considered a law from the perspective of that agent. A sentence is universal iff it begins with a universal quantifier. We shall also speak of generalizations in order to refer to the laws of an epistemic state. We follow the convention that generalizations may be represented by implications in propositional logic since propositional variables may stand for events at the type-level. We do not distinguish between laws and generalizations. Obviously, our syntactic characterization of laws is wider than the intuitive notion of law of nature. We shall say more about the latter notion in the next chapter.

Likewise, we work with a syntactic characterization of beliefs concerning presumed atomic facts: such beliefs are expressed by an atomic sentence or the negation of such a sentence. For simplicity, we speak of facts when referring to beliefs about presumed atomic facts. The notion of fact is always considered relative to an epistemic perspective here.

The syntactic characterization of laws may be thought to run into the following problem. Suppose I believe A, which stands for the occurrence of an event. By classical logic, I also believe $A \vee B$, and consequently $\neg A \rightarrow B$, for any arbitrary sentence B. The important point to note here is that a belief base contains only explicit beliefs. Put more carefully, belief bases are intended to represent the explicit beliefs of an epistemic state. With this qualification in mind, we can say that I explicitly believe A, but not so $\neg A \rightarrow B$. The latter sentence is therefore not a member of the belief base which is supposed to model my epistemic state.

Using belief revision theory, we can furthermore explain the distinction between law-like and trivial implications as follows. Suppose we believe $A \rightarrow B$ in an epistemic state S. In formal terms, $A \rightarrow B \in K(S)$. Then we can say that $A \rightarrow B$ is a non-trivial belief in S iff $A \rightarrow B \in K(S) \div \neg A$ and $A \rightarrow B \in K(S) \div B$. That is, an implication is non-trivially believed iff we continue to believe it if our beliefs are contracted by the negation of the antecedent or the consequent of that implication. Believing an implication non-trivially means that we believe in a connection between antecedent and consequent. If, by contrast, we were to believe $A \rightarrow B$ merely because we believe $\neg A$, then we would have to give up $A \rightarrow B$ when giving up $\neg A$.

Similar considerations apply to the syntactic characterization of laws in terms of first-order sentences which begin with a universal quantifier. Suppose I believe A, which does not have any occurrences of any quantifier. Then I also believe $\forall x A$. But the latter is not an explicit belief of mine. I simply do not explicitly believe that it holds true for all objects x that Aristotle was a philosopher. Put differently, a non-trivial universal sentence must have non-trivial instances. Such instances may be obtained by replacing some occurrence of a variable with a constant using the inference rule of Universal Elimination in a system of natural deduction. In yet other words, a sentence which begins with a universal quantifier is considered a law only if the universal quantifier occurs non-vacuously at the beginning of that sentence.

If we accept a law, we accept some universal sentence or an implication whose propositional variables have a type-level meaning. Adopting a law, however, does not necessarily imply that the agent believes all of its instances to be true. This is not so for two reasons. First, we can take the belief base as an idealized theory which captures the phenomena of some domain quite successfully, while there remain cases in which some instance of a law fails to hold true. Second, belief revision theory allows for a distinction between strict and non-strict laws. It even contains an account of default and *ceteris paribus* reasoning. We take the notion of non-strict law to comprise both default and *ceteris paribus* laws. The two types of laws may well overlap substantially.

Thus we have explained the basic concepts of belief revision theory, including the notion of a belief base with a ranking of epistemic priority. These explanations suffice for a general account of epochetic conditionals in belief revision theory. Further details concerning the operations of an expansion, contraction, and revision of a belief base are explained in Appendix B.

4 Epochetic Conditionals

Our epochetic conditionals draw on the Ramsey Test, an epistemic evaluation recipe for conditionals devised by Ramsey (1931a). Its core idea has been pointedly expressed by Stalnaker (1968, p. 102):

> First, add the antecedent (hypothetically) to your stock of beliefs; second, make whatever adjustments are required to maintain consistency (without modifying the hypothetical belief in the antecedent); finally, consider whether or not the consequent is then true.

It was then Gärdenfors (1978) who translated this test into the language of belief changes and who insisted more forcefully than Stalnaker on an epistemic understanding of conditionals. Using the AGM framework, he was able to define a semantics of conditionals in terms of belief changes:

$$A > C \in K(S) \text{ iff } C \in K(S) * A \tag{RT}$$

where $>$ stands for the conditional connective. $K(S) * A$ designates the revision of the beliefs of an epistemic state S with the sentence A. The Ramsey Test thus defines that a conditional $A > C$ is to be accepted in a belief system $K(S)$ iff the consequent C is in the revision of $K(S)$ by the antecedent A. Unlike Gärdenfors (1978), we require that A and C be non-conditional sentences.

We now define a conditional with the following intuitive meaning: $A \gg C$ iff, after suspending any beliefs in $K(S)$ as to whether A and C are true or false, we can infer C from A in the context of the remaining beliefs. In more formal terms:

Definition 12. Belief function $B(A)$
Let A be a sentence and S an epistemic state.

$$B(A) = \begin{cases} A & \text{if } A \in K(S) \\ \neg A & \text{if } \neg A \in K(S) \\ \bot & \text{otherwise.} \end{cases}$$

$$A \gg C \in K_{>}(S) \text{ iff } C \in (K(S) \div B(A) \vee B(C)) + A. \tag{SRT}$$

Equivalently,

$$A \gg C \in K_{>}(S) \text{ iff } (K(S) \div B(A) \vee B(C)), A \vdash C$$

where $\vdash$ designates the relation of provability in classical logic. The first step of (SRT) consists in an *agnostic move* which lets us suspend judgement about the antecedent and the consequent. The contraction by $B(A) \vee B(C)$ gives us an epistemic state in which we do not believe A, B, $\neg A$, or $\neg B$. For example, if we believe the antecedent A to be true in the epistemic state S, we need to contract the beliefs of S by A. If, by contrast, we believe A to be false, we need to contract by $\neg A$. If we believe neither A nor $\neg A$, no actual contraction is needed. This is expressed by a contraction by $\bot$, which stands for falsum or a contradiction. Provided the beliefs in S are consistent, contraction by $\bot$ does not change the beliefs in S.[8]

Once we have suspended judgement about antecedent and consequent, we check whether or not we can infer the consequent C from the antecedent A in the context of the remaining beliefs of the epistemic state. If so, $A \gg C \in K_{>}(S)$. Otherwise, $A \gg C \notin K_{>}(S)$. $K_{>}(S)$ is the belief set of the epistemic state S, extended by the Ramsey Test for some conditional.

[8] Our conditional $\gg$ is inspired by a proposal by Rott (1986) to strengthen the Ramsey Test.

We have thus defined an epochetic conditional for epistemic states S in general to indicate that such a conditional is not necessarily tied to belief bases. We assume henceforth, however, that the respective epistemic state S is given by a ranked belief base $\mathbf{H} = \langle H_1, \ldots, H_n \rangle$. Thus we obtain:

$$A \gg C \in K_{>}(\mathbf{H}) \text{ iff } C \in (K(\mathbf{H}) \div B(A) \vee B(C)) + A.$$

This is the definition of our strengthened Ramsey Test upon which we build our Humean analysis of causation. $K(\mathbf{H})$ is given by the classical logical closure of the union of all components of $\mathbf{H}$: we believe a sentence A in an epistemic state $S = \mathbf{H}$ iff A is a classical consequence of the set $H_1 \cup \ldots \cup H_n$ of all beliefs of the ranked belief base $\mathbf{H}$.

Recall that the contraction operation $\div$ is constrained by the following principle: when contracting an epistemic state by a belief A, retain as many of the currently held beliefs as possible—without however retaining A. Belief changes are guided by the maxim of minimal mutilation, which goes back to Quine (1961). We explain further details and give fully explicit definitions of the contraction and revision of a ranked belief base in Appendix B.

5 Causation

In the previous section, we have worked out the semantics of an epochetic conditional $\gg$ in belief revision theory. This conditional has been defined for epistemic states which have the logical form of a ranked belief base with two levels: a set of laws and a set of presumed facts. The laws are represented by implications and universal sentences, the presumed facts by literals. Unlike causal models of the form $\langle M, V \rangle$ in Part I, no modal or causal notions are assumed for the formulation of laws and facts. Nor do we assume any more substantial notion of law of nature, which would go beyond the syntactic characterization of laws in the above section. We will further develop our syntactic approach to laws so as to capture some proper notion of law of nature in the next chapter when addressing the problem of spurious causation.

Our new epochetic conditional $\gg$ gives us an approximate explanation of what it is for a proposition to be a reason for another.

Definition 13. Reason
A is a reason for B—relative to an epistemic state S—iff A and B are in $K(S)$, and $A \gg B$ is in $K_{>}(S)$.

Readers who skipped the more technical section on belief revision theory may note that $K(S)$ stands for the beliefs of the epistemic state S, including beliefs implied by the explicit beliefs of S. $K_{>}(S)$ contains all epochetic conditionals accepted from the epistemic perspective of S. It is defined by a strengthened Ramsey Test. Epistemic states are given by a finite set of explicit beliefs together with a simple ranking of priority among these beliefs. Basically, laws have epistemic priority over presumed facts.

Note that we use the symbol $\gg$ for a variety of different epochetic conditionals. In Part I, various such conditionals have been defined for causal models $\langle M, V\rangle$. In Part II, $\gg$ is defined for ranked belief bases whose members do not contain any explicitly causal notions. We will further refine and strengthen this conditional in what follows. We trust the reader will gather the respective meaning of $\gg$ from the context.

How do we get from reasons to causes? We have argued that the Humean convention is worth a reconsideration in order to discriminate between mere inferential reasons and reasons which are causes or correspond to causes. For now, the convention leads us to the following analysis:

Definition 14. Cause
Let C and E be events. C is a cause of E—relative to an epistemic state S—iff

(C1) $C, E \in K(S)$,

(C2) $C \gg E \in K_{>}(S)$, and

(C3) C precedes E.

As in Part I, we use upper case Latin letters for both events and propositions which claim that the respective event occurs.

How should we understand the temporal order between two events which partly overlap with one another? When we say that C precedes E, we mean that C comes into being before E. Suppose the temporal existence of C or E, or both, is represented by an interval. Recall that the left-hand limit of a temporal interval represents the beginning of it. The left-hand limit of a time point t may be identified with t itself. Then we say that C precedes E iff the left-hand limit of C is temporally earlier than the left-hand limit of E.

This analysis of causation is still very much a preliminary one. As explained at the beginning of this section, there remain the problems of spurious, simultaneous, and backward causation. We will address and solve these problems in the subsequent chapters. We must also wonder which of the scenarios of actual causation, discussed at length in Part I, are captured by our preliminary Humean analysis. Now, it is easy to show that causal scenarios of overdetermination, conjunctive scenarios, and combinations of the two are not a problem. Some scenarios of prevention and double prevention are captured as well.[9]

To capture scenarios of preemption, we may utilize the notion of active path in a manner analogous to our analysis in Part I. Recall that this notion rests on the notion of inferential dependence on the candidate cause, which in turn has been explained in terms of natural deduction. Preempted causes fail to have an active path to the effect since any deduction of the effect requires reasoning by cases in such a manner that one subproof contains inferential steps which do not depend on the candidate cause. This observation seems to hold as well for at least some first-order representations of preemption scenarios without causal models. There remain some details to be worked out, though.

These observations give rise to a more general question: could we translate our epochetic analysis in Part I into one which is based on the Humean convention instead of structural equations? We tend to answer the question in the affirmative. The main challenge of the translation concerns condition (3) of the epochetic conditional $\gg$ for causal models, defined in Chapter 2. This condition protects structural equations which determine variables which are descendants of the presumed cause. This poses a problem for the translation since the concept of descendant rests on the causal graph of the respective scenario. It is not obvious how we can derive this graph from our reductive analysis in Part II. To translate the condition in question, we may require that we me must not retract laws whose descriptive terms stand for events which are simultaneous or later than the presumed cause. Again, this is a mere outline of a translations which requires further elaboration.

The envisioned translation would have important benefits. Most impor-

[9] See Andreas and Günther (2020) for details and a demonstration of these results.

tantly, it would give us a reductive analysis of causation which captures an unprecedentedly large range of causal scenarios. However, our simple Humean analysis is already awaiting further refinements in view of the problems of spurious, simultaneous, and backward causation. If we were to merge the latter refinements with a translation of our final analysis in Part I, the resulting analysis may lose appeal, simply because of its complexity. Also, we are hesitant to make the reader study once again a large number of causal scenarios with us.

With these concerns in mind, we decided to go for another strategy. We will focus in Part II on the problems spurious, simultaneous, and backward causation. Then we say what it is for a causal model to be verified by an epistemic state which contains only laws and facts without any explicitly causal notions. The idea is, roughly, that a causal model $\langle M, V\rangle$ is verified by an epistemic state S iff all elementary causal relations of the causal model analysis come out as genuine causal relations on the reductive analysis, and vice versa. This biconditional continues to hold if we change the valuation from V to V' and revise S by V'. A causal relation between two literals L_A and L_B is elementary in a causal model $\langle M, V\rangle$ iff there is a directed edge from A to B in the causal graph of M and L_A is a cause of L_B relative to $\langle M, V\rangle$.

The relation of verification between an epistemic state and a causal model is understood in analogy to the model-theoretic relation of verification. In model-theoretic semantics, we say that a certain interpretation, or structure, of a formal language verifies a set Γ of sentences iff all members of Γ are true on this interpretation. One and the same set of sentences may be verified by a variety of different interpretations. Likewise, a causal model may be verified by a variety of different epistemic states.

The architecture of our theory is thus as follows. At the most fundamental level, we have epistemic states whose explicit beliefs are free of explicitly causal notions. But we do have information about temporal relations among events at this level. At the second level, we define causation using belief revision theory, the Humean convention, and some refinements concerning the inferential relations of the epochetic conditional $\gg$. Then, at the third level, we have causal models which may be verified by epistemic states of the first level, given the reductive analysis. Finally, at the fourth level, we define actual causation in terms of causal models, as done in Part I.

On the face of it, it may seem surprising to use epistemic states as foundation of causal models. Note, however, that Hume's project is to analyse causation in terms of concepts which are less mysterious and less theoretical than causation is. To make progress with this project, it would not be very helpful to invoke some metaphysical notion of reality at an early stage. We will not forget about the world, though. In the Conclusion and Synthesis of this book, we will outline how our theory could be anchored in reality.

Chapter 8

Spurious Causation

Spurious causal relations may well be considered the most severe problem for a Humean approach to causation. There are causal scenarios where we have a regular connection between two types of events, one event always precedes the other, and yet we do not consider the earlier event a cause of the later event. This happens in common cause scenarios when the two effects are not simultaneous. Not surprisingly, the problem of spurious causation does also arise for our Humean analysis set forth in the previous chapter. This calls for an inferential characterization of spurious causal relations.

In this chapter, we show that spurious causes differ from genuine causes in terms of the laws used on the inferential path to the effect. In essence, there is no forward-directed inferential path from the spurious cause to its effects such that all laws on this path are non-redundant. Put differently, the laws on the inferential path from the genuine cause to its effect have a greater unificatory power than the laws on the inferential path from the spurious cause to this effect. The notions of redundancy and unificatory power will be explicated relative to the beliefs of an epistemic state.

1 Common Causes

Let us a begin with a simple and abstract scenario of spurious causation. Suppose event C is a common cause of the events A and E, as depicted by Figure 62. Further, suppose A can be caused only by C-type events, while C-type events cannot fail to bring about an E-type event. C always produces A. The causal relations between the common cause and its effects are forward-directed in time. Let us finally assume that event A precedes event E.

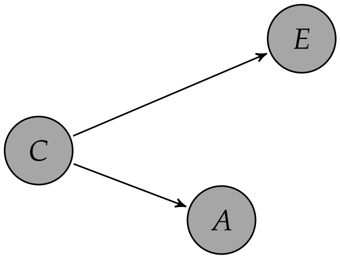


Figure 62: Common cause

It is easy to see why the scenario in question is trouble for an inferential analysis of causation which takes a simple Humean approach to the direction of causation. Since A-type events can only be caused by C-type events, we can infer C from A. Since C cannot fail to cause E, we can infer E from C. Given the inference relation is transitive, we can therefore infer E from A. Together with A preceding E, this implies that A is a cause of E. This verdict, however, does not agree with our causal judgements. We think of A as a mere spurious cause of E, as will become more obvious when we look at concrete scenarios of common causes below.

Our presumed knowledge about the abstract causal scenario may be represented by the following belief base:

$C \leftrightarrow A,\ C \rightarrow E$
C, A, E

For simplicity, we leave beliefs about the temporal relations among A, C, and E implicit. Recall that the different levels of this belief base stand for different epistemic priorities: the implications have higher priority than the atomic sentences. Given this representation, we can solve the problem of spurious causation by an additional constraint on the inferential paths from the antecedent to the consequent of our conditional $\gg$:

> $A \gg B$ if and only if, after suspending judgment about A and B, we can infer B from the supposition of A such that each inferential step is forward-directed in time.

We say that an inferential step is *forward-directed in time* iff no premise asserts the occurrence of an event which precedes the event asserted by the conclusion. The notion of forward-directed inferential step is thus understood in the weak sense of not being backward-directed in time. To give a simple example, the inference from C to A is forward-directed in time, while that from A to C is backward-directed in time.

There remain some ambiguities, though. For example, not every inferential step proceeds from premises which assert the occurrence of an event. The following definition addresses this and related ambiguities.

Definition 15. $H \vdash_F C$
Let H be a set of sentences and C be a sentence. Only literals and conjunctions of literals are taken to assert the occurrence or absence of an event. We say there is a *forward-directed deduction* of C from H—in symbols $H \vdash_F C$—iff there is a natural deduction of C from H such that, for all inferential steps P/I, if I asserts the occurrence or absence of an event, then this event or absence does not precede any event or absence asserted by a premise in P or by a premise in a subproof which is a member of P. This requirement applies to all inferential steps in the main proof and any subproof.

The refinement of our epochetic conditional falls now into place:

$$A \gg_F C \in K_{>}(S) \text{ iff } \bigcup(S \div B(A) \vee B(C)), A \vdash_F C. \tag{SRT$_F$}$$

Recall that S stands for an epistemic state which has the form of a belief base with different levels of epistemic priority. $\bigcup(S \div B(A) \vee B(C))$ is the set of explicit beliefs after suspension of judgement on A and B.

Let the epistemic state S now be given by the above belief base $\langle\{C \leftrightarrow A, C \rightarrow E\}, \{A, C, E\}\rangle$. For simplicity, we leave the temporal relations among the events implicit. It is then easy to show that $A \gg_F E$ does not hold. Hence, if we adopt (SRT$_F$) instead of (SRT) (as defined in Section 2), we obtain the favourable result that A is not a cause of E, as it should be. The notion of forward-directed deduction may be considered a proof-theoretic variant of the Humean convention.

However, the suggested approach to spurious causation does not go very far. The problem is that, for the scenario in question, it holds universally true that an A-type event is followed by an E-type event. Hence, our belief system may well contain the implication $A \rightarrow E$, which says that an E-type event occurs whenever an A-type event does. There then is a forward-directed inferential path from A to E. Hence, even our refined analysis wrongly counts A as a cause of E once $A \rightarrow E$ is adopted as an explicit belief.

To make further progress, let us study some concrete scenarios of spurious causation. The drop of a barometer, which is followed by a storm, is a classical example. When a low-pressure system approaches a specific region, a barometer located in this region drops. The air pressure indicated by the barometer falls. Low-pressure areas often bring stormy weather. In this scenario, we have two properly causal relations and one spurious causal relation. The arrival of a low-pressure system causes the barometer to drop and brings stormy weather. It is a common cause of two different effects: the barometric column falls and strong winds occur. Since the barometer starts falling before the wind becomes actually stronger, it holds true that strong winds follow the drop of the barometer. However, we do not consider this drop a genuine cause of the storm. Thinking that stormy weather is caused by a falling barometer strikes us as incorrect.

The barometer example is often discussed in the context of probabilistic causal relations (see, e.g., Reichenbach (1956, p. 193)). Rightly so. There is no deterministic correlation between low-pressure systems and stormy weather. Meteorologists are cautious to not overestimate the predictive power of barometer readings.[1] Since our investigation is confined to deterministic causation, the barometer example is not well suited to study spurious causal relations.

Here is a better example. As is well known, water freezes at temperatures below zero degrees Celsius. By convention, the zero point of the Celsius scale equals the freezing point of water. At the same time, we consider air temperatures below zero degrees Celsius a cause of water to freeze. And

[1] 'Recall that anticyclones (high-pressure cells) are associated with clear skies and that cyclones (low-pressure cells) frequently bring clouds and precipitation. Thus, by noting whether the barometer is rising, falling, or steady, we have *some indication* of the forthcoming weather.' (Lutgens et al. (2013, p. 253), emphasis added). See also Lutgens et al. (2013, p. 243n).

we consider the air temperature a cause of the thermometer reading. This is particularly obvious for liquid thermometers which are based on the correlation between the volume of the liquid in the thermometer and the temperature of the surrounding air, but it holds also true for other types of thermometers. Hence, air temperature causes a certain thermometer reading, just as air pressure causes a certain barometer reading. In this simple causal scenario, we can recognize two genuine causal relations:

(1) If the air temperature drops below zero degrees Celsius, liquid water starts to freeze.

(2) If the air temperature drops below zero degrees Celsius, a thermometer indicates an air temperature below zero degrees Celsius.

The following causal relation, however, is of the spurious type:

(3) If a thermometer indicates an air temperature below zero degrees Celsius, liquid water starts to freeze.

We view only the first two conditionals as properly causal in the sense that the antecedent stands for a cause and the consequent for a corresponding effect. The third conditional represents a spurious causal relation. However, all conditionals satisfy the Humean convention: the antecedent precedes the consequent. It takes some time for water to freeze once air temperature drops below zero degrees Celsius. No thermometer immediately responds to a change of the temperature in the environment. Finally, a thermometer drops below zero degrees before we can observe the water to freeze. In the background we have processes in which a thermodynamic system approaches thermodynamic equilibrium. Such processes take some time. The spurious cause occurs before the other effect of the common cause because a thermometer approaches thermodynamic equilibrium faster than a medium-sized reservoir of water.

For clarification, we should point out that the conditionals (1), (2), and (3) have exceptions. They must be read with a *ceteris paribus* qualification or taken as elements of an idealized model. Big reservoirs of water, such as lakes, start to freeze only if the temperature drops below zero degrees for several days, if not weeks. So we assume that the water reservoir has medium size and that the temperature drops significantly below zero degrees for several days. Also, we implicitly assume that there is no salt in the water that would change the freezing point. The thermometer must be properly gauged. Such qualifications, however, may not prevent us from using deterministic models. After all, we study causal judgements in deterministic models about the world. Sets of *ceteris paribus* laws give rise to such models. For simplicity, we often leave certain conditions of properly applying deterministic models implicit. In the present example, we take the conditionals as strict laws of a simplified theory about the aggregate phase of water in everyday contexts. The theory intentionally abstracts from kinetic theories of heat, which however will be considered in Section 12.

There are furthermore real-world examples of spurious causation which do not involve measurement devices. A causal analysis of lightning should distinguish at least three events: (i) an electrostatic discharge through the atmosphere between a cloud and the ground, (ii) the flash of the lightning, and (iii) the thunder. Physics tells us that the electrostatic discharge is—via the rapid production of heat within the region of the air where electricity is conducted—the common cause of the flash and the thunder. From the viewpoint of physics, the flash is not a genuine cause of the thunder. It's a mere spurious cause of the latter.

Why are we hesitant to call the flash in the sky a cause of the thunder? Why is a temperature reading below zero degrees Celsius not considered a proper cause of liquid water to freeze? More generally, why are measurement readings considered mere spurious causes of phenomena which we think are actually caused by the measured quantities? After all, we have a rather strict correlation between spurious causes and their effects. Also, the spurious cause precedes its effect. What is wrong with spurious causes?

To answer this question, we take recourse to unificationist ideas about explanation: the correlations between spurious causes and their effects is less important to account for diverse phenomena in a unified manner than the laws used to infer effects from genuine causes. For example, general correlations between measurement readings and natural phenomena are derivable from more fundamental laws about these phenomena. The correlation between a thermometer reading below zero degrees and the freezing of liquid water can be derived from our knowledge about the working of thermometers and the physical properties of water. Hence, there is no need to view such a correlation as a proper law of nature. Put differently, laws directly connecting spurious causes with their effects are redundant, while laws used to infer effects from genuine causes have some unificatory power.[2] These observations suggest an inferential characterization of spurious causation.

Definition 16. Spurious Causation
C is a mere spurious cause of E iff all forward-directed inferential paths from C to E use some redundant law.

If, by contrast, C is a genuine cause of E, we have a forward-directed inferential path from C to E such that all laws used on this path are non-redundant. Such laws are key to causal inferential paths. Of course, we need to say more about the notions of non-redundancy and unification.

Our approach to spurious causation may be motivated by the best system account of laws of nature, commonly referred to as the *Mill-Ramsey-Lewis* account of laws. For a genuine cause, there is a forward-directed inferential path to the effect such that all laws used on this path are laws of nature—in the sense of at least some construals of the best system account. Mere spurious causes lack this property.

A clarification concerning the concepts of law, generalization, and implication is in order here. Recall that we adopted a minimalist and syntactic understanding of the notion of law in the previous chapter: any universal sentence and any implication explicitly believed to be true—by the respective agent—is considered a law by that agent. Thanks to this minimalist notion of law, we can represent redundant and non-redundant laws in the setting of classical logic. Moreover, we follow the convention that laws may be represented by implications in propositional logic since propositional constants may stand for events at the type-level. We do not distinguish between laws and generalizations. A law of nature in the narrow sense is called a *proper law* or simply a *law of nature*.

[2] The unificationist account of explanation goes back to Friedman (1974) and Kitcher (1989). We will say more about this account in Section 12.

2 Non-Redundant Laws

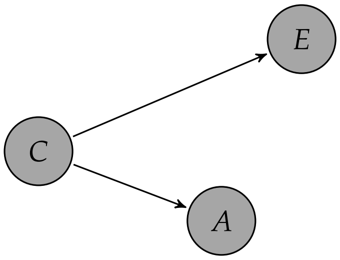
What is a unificatory law? Let us begin with the distinction between redundant and non-redundant sentences. Here is a simple proposal: ϕ is redundant in a set Γ of sentences iff it can be inferred from the other members of Γ. In symbolic notation:

Definition 17. Redundancy
A sentence ϕ is redundant in a set Γ of sentences—relative to an inference relation $\vdash$—iff $\phi \in \Gamma$ and $\Gamma \setminus \{\phi\} \vdash \phi$.

Put simply, ϕ is redundant in Γ iff, when taking ϕ out of Γ, we can get it back from the rest of Γ using inferences of the respective logic. This understanding of a redundant sentence gives us a straightforward notion of a non-redundant one:

Definition 18. Non-redundancy
A sentence ϕ is non-redundant in a set Γ of sentences—relative to an inference relation $\vdash$—iff $\phi \in \Gamma$ and $\Gamma \setminus \{\phi\} \nvdash \phi$.

Put simply, ϕ is non-redundant in Γ iff, once we take it out of Γ, we cannot get it back by logical inferences. Of course, we need to specify the logic according to which we understand the notion of redundancy. In the absence of further results, classical logic suggests itself as the simplest choice. Let us now formalize some of the above scenarios of spurious causation to see if the proposal works with classical logic. In the previous section, we began with an abstract scenario of spurious causation, which was formally represented as follows:



$C \leftrightarrow A, \quad C \to E, \quad A \to E$
$C, \quad A, \quad E$

Figure 62: Common cause

Now, the key idea of our envisioned analysis is that a candidate cause is a mere spurious cause of E iff we need to use some redundant law in order to infer the effect in a forward-directed manner. There is no inferential path from the spurious cause to the effect such that all laws used on this path are non-redundant. For the present scenario, our envisioned analysis implies that $A \rightarrow E$ is redundant in the set $H = \{C \leftrightarrow A, C \rightarrow E, A \rightarrow E\}$ relative to classical logic. This holds true indeed. The bad news, however, is that $C \rightarrow E$ is likewise redundant in the set H relative to classical logic. For this to be seen, start a subproof with the assumption of C. Then you can infer E using $C \leftrightarrow A$ and $A \rightarrow E$. By Implication Introduction, you can infer $C \rightarrow E$. Our envisioned analysis therefore implies that C is a spurious cause of E, which is absurd.

However, let's not give up too quickly on our inferential approach to spurious causation. Our proposal may still work for concrete scenarios of spurious causation for which further information about the causal factors of the different effects is available. The present formalization of a common cause may turn out too abstract to allow for a proper distinction between spurious and genuine causes. Also, it's worth considering non-classical inference relations in order to understand the distinction between redundant and non-redundant laws. When studying concrete scenarios of spurious causation, we will see below that classical logic does not always suffice to infer the common cause from the spurious cause in the respective causal scenario. This inferential step, however, is needed in order to show that the law which directly connects the spurious cause with its effect is redundant. Let us therefore study abductive inferences in the next section.

3 Abductive Inferences

The notion of an abductive inference originates from the work of Charles Sanders Peirce. It is introduced there as an inference to a hypothesis that is capable of explaining certain facts. Such an inference has the following form:

> The surprising fact, C, is observed.
> But if A were true, C would be a matter of course.
> Hence, there is reason to suspect that A is true.
> (Peirce 1931, Vol. 5, p. 189)

If we translate the condition ‘if A were true, C would be a matter of course’ with ‘C is inferable from A’, we obtain the following inference rule:

$$\frac{C \qquad \begin{array}{c} A \\ \vdots \\ C \end{array}}{A}.$$

In essence, abductive inferences are inverted deductions. This view is fundamental to logical approaches to abductive reasoning. $A \dots C$ stands for a subproof in which C is derived from A.[3]

Logical approaches to abductive reasoning hold on to the view that abduction is an inference to a hypothesis which explains certain phenomena, which have been observed. If we use abductive reasoning to establish causally explanatory hypotheses, we infer causes from given effects. An account of abductive reasoning may thus prove useful to capture the inference from the spurious cause to the common cause in common cause scenarios. We need to establish this inference in order to show that the law which directly connects the spurious cause with its effect is redundant relative to a certain logic. Of course, we cannot simply introduce an inference rule of the form

$$\frac{C \text{ causes } E, \quad E}{C}.$$

This would be flatly circular since we would have to take some causal relations as primitively given. However, we have already established a proto-theory of causation which does not take any causal or modal notions as antecedently available. According to this proto-theory, a central requirement for causation is that there is a forward-directed inferential path from cause to effect. Our proto-theory thus suggests the following abductive inference rule:

[3] Our inference rule is inspired by the notion of a conjectural inference relation in Flach (2000, p. 96). See Schurz (2008) for an overview of different types of abductive inferences. Brewka et al. (1997, Ch. 5) survey accounts of abductive reasoning in knowledge representation.

$$\dfrac{E \qquad \begin{matrix} C \\ \vdots_F \\ E \end{matrix}}{C}\;.$$

The index F in the symbol $\vdots_F$ indicates that all inferential steps in the subproof from C to E are forward-directed in time—in the weak sense that none is backward-directed in time. If an inferential step consists of sentences which are not about events, then this inferential step is said to be forward-directed as well. We can say it is vacuously forward-directed.

There are still some problems with this proposal, though. First, suppose in a disjunctive scenario—where an event may be brought about by different causes—the effect event occurs. The present inference rule would then allow us to infer that all disjunctive causes are occurrent. But this conclusion may not be true, even if we assume that everything happens for a reason or cause. We should only infer that at least one of the disjunctive causes occurs.

A second problem concerns the logical form of the assumption and conclusion of the subproof. The notation implicitly assumes that both are atomic sentences. This constraint, however, excludes conjunctive causes. On the other hand, if we do not impose further constraints on assumption and conclusion of the subproof, we end up with a far too liberal inference rule of abduction. For example, we could infer $A \wedge B$ from B if A stands for an event which does not occur later than event B. This would be absurd. In light of these problems, we suggest the following inference rule of abductive reasoning:

$$\dfrac{\beta \qquad \begin{matrix} \alpha_1 \\ \vdots_F \\ \beta \end{matrix} \qquad \begin{matrix} \\ \ldots \\ \\ \end{matrix} \qquad \begin{matrix} \alpha_n \\ \vdots_F \\ \beta \end{matrix}}{\alpha_1 \vee \ldots \vee \alpha_n}\;. \qquad \text{(Abduction)}$$

If we can infer β from each of α_1, α_2, and α_n in a forward-directed manner, and β is given, then this rule allows us to infer $\alpha_1 \vee \ldots \vee \alpha_n$. However, we need to take the following conditions of application into account:

(A1) For all i $(1 \leq i \leq n)$, $\alpha_i \nvdash_{Cl} \beta$.

(A2) If there is a forward-directed subproof from ϕ to β, then there is some i $(1 \leq i \leq n)$ such that $\phi \vdash_{Cl} \alpha_i$.

(A3) For all i $(1 \leq i \leq n)$, there is no ϕ such that there is a forward-directed subproof from ϕ to β, $\alpha_i \vdash_{Cl} \phi$, and $\phi \nvdash_{Cl} \alpha_i$.

These conditions prevent us from inferring 'too much' using the inference rule Abduction. Condition (A1) says that β must not be a logical consequence of any assumption $\alpha_1, \ldots, \alpha_n$. This condition blocks the inference from B to $A \wedge B$, which we should not consider valid.

We may wonder how we can infer β from α_i if the former is not a logical consequence of the latter. To see this, recall that a subproof from α_i to β may well depend on premises and conclusions in the superordinate proof.

The symbol $\vdots$ stands for a mere subproof and may not be confused with the relation of derivability, often abbreviated by the symbol $\vdash$. Laws of background theories about the respective causal scenario may be given in the premises of the superordinate proof.

Condition (A2) requires us to consider all forward-directed subproofs of a given sentence β when inferring a cause of why β is true by Abduction—except for those subproofs whose assumption is logically at least as strong as the assumption of a subproof which has already been considered. We say that a subproof from ϕ to β has been considered if such a subproof figures as antecedent of the abductive inference in question. A few examples may help us understand Condition (A2) and its motivation.

Condition (A2) solves the above problem of disjunctive causal scenarios. Suppose A and B are individually sufficient to bring about E. We know that E occurs, and we have forward-directed subproofs from, respectively, A and B to E. Then (A2) ensures that we cannot use the abductive inference rule to infer A. Nor can we infer B. For in the following abductive inference from E to B, we have ignored the subproof from A to E:

$$\frac{E \qquad \begin{array}{c} B \\ \vdots_F \\ E \end{array}}{B}\ .$$

Not considering the subproof from A to E in this inference amounts to a violation of (A2). We are only allowed to infer $A \vee B$, as it should be:

$$\frac{E \qquad \begin{matrix} A \\ \vdots_F \\ E \end{matrix} \qquad \begin{matrix} B \\ \vdots_F \\ E \end{matrix}}{A \vee B}\,.$$

Unlike the former, the latter inference satisfies condition (A2).

Condition (A3) demands that each α_i is among the (classically) logically weakest assumptions from which we can infer the event β in a forward-directed manner. In other words, for each α_i, there must not be ϕ such that ϕ is (classically) logically weaker than α_i, and yet we can infer β from ϕ in a subproof in a forward-directed manner. Condition (A3) prevents us from abductively inferring a conjunction of events when just one conjunct is sufficient to bring about the effect. Suppose there is a complex scenario with several events, while C is the only cause of E. So there is a corresponding forward-directed subproof from C to E. E cannot be inferred in a forward-directed manner from events other than C, given our theory about the causal scenario. We know that E occurs, but we do not know which other events occur. Obviously, we can infer from E that C occurs by Abduction. But we cannot infer, for example, $D \wedge C$, even if D precedes E. The following application of the abductive inference rule violates condition (A3):

$$\frac{E \qquad \begin{matrix} D \wedge C \\ \vdots_F \\ E \end{matrix}}{D \wedge C}\,.$$

Condition (A3) is violated here because C is logically weaker than $D \wedge C$ and E can be inferred from C in a subproof in a forward-directed manner.

Let us also look at a conjunctive scenario. Suppose there is a forward-directed subproof from $A \wedge B$ to E. E cannot be inferred in a forward-directed manner from events other than the joint occurrences of A and B. The application of the abductive inference rule is thus straightforward:

$$\dfrac{E \qquad \begin{array}{c} A \land B \\ \vdots_F \\ E \end{array}}{A \land B}\ .$$

This inference satisfies Condition (A2). Note that this condition does not require us to consider the (forward-directed) subproof from $B \land A$ to E if we already considered the one from $A \land B$ to E. For, obviously, the assumption of the former is logically equivalent to the assumption of the latter.

Does our abductive inference rule lead to problems in causal scenarios of indirect causation? Suppose A brings about B, which in turn brings about E:

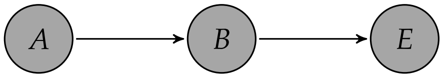


Figure 63: Causal chain

So there is a forward-directed subproof from A to B and one from B to E. E cannot be inferred (in a forward-directed manner) from an event other than B, and B cannot be inferred (in a forward-directed manner) from an event other than A. We know that E occurs, but have no explicit knowledge about the occurrences of other events. Of course, we want to infer B from E by Abduction, and then A from B, again by Abduction. However, the following application of the inference rule is not correct:

$$\dfrac{E \qquad \begin{array}{c} B \\ \vdots_F \\ E \end{array}}{B}\ .$$

This inference violates condition (A2). For there is a (forward-directed) subproof from A to E, and A is neither logically stronger than nor logically equivalent to B. This subproof has not been considered in the present inference. Hence, we have to be content with the following abductive inference:

$$\frac{E \qquad \begin{matrix} A \\ \vdots_F \\ E \end{matrix} \qquad \begin{matrix} B \\ \vdots_F \\ E \end{matrix}}{A \vee B} \; .$$

This type of problem, however, has already been taken care of by the inference rules of classical logic. Note that we can infer, by classical logic, B from A and the law saying that event B follows event A. Trivially, we can infer B from B. Reasoning by cases—also referred to as *Disjunction Elimination*—therefore allows us to infer B from $A \vee B$. Then we can infer A from B by Abduction. Thus we have inferred both A and B from E, as it should be.

Two objections are worth considering. Obviously, conditions (A1) to (A3) are crucial to understanding and applying the proposed abductive inference rule. These conditions are presented in a semi-formal fashion. It is an explanation additional to the purely formal presentation of the abductive inference rule. This should not be considered a problem. Note that even some natural deduction inference rules of classical logic rest on additional explanations, which are not fully formalized. A case in point are explanations as two which constants can be used in a subproof when applying the inference rule Existential Elimination. Some rules determining the proper use of a calculus always escape complete formalization. This is an important lesson from Wittgenstein's *Philosophical Investigations* (1953).

Another objection may target the fact that conditions (A1) to (A3) rest on propositions about derivability in classical logic. Admittedly, this is not desirable since it makes the envisioned system of abductive reasoning impure. The system is not only built on top of the inference rules of classical logic, but does also rest on metalogical propositions concerning derivability in classical logic. However, the increased expressive power of abductive reasoning comes at a price. Certain nonmonotonic logics have inference rules whose conditions of application rest on propositions about derivability and entailment in classical logic as well. A case in point is Reiter's default logic (see Antoniou (1997, Part III)). Likewise, our account of default reasoning in the previous chapter is based on some metalogical concepts of classical logic. Specifically, we need the concepts of consistency and entailment in classical logic for this account.

We are now in a position to define an inference relation of abductive reasoning on top of classical logic:

Definition 19. $\Gamma \vdash_a \phi$
Let $\vdash_{Cl}$ be a natural deduction system of classical first-order logic. Let Γ be a set of first-order sentences and let ϕ be such a sentence. We say that ϕ is inferable from Γ by abductive reasoning—and write $\Gamma \vdash_a \phi$—if and only if there is a natural deduction derivation of ϕ from Γ using the inference rules of $\vdash_{Cl}$ and the inference rule Abduction.

We must wonder whether this inference system of abductive reasoning constitutes a proper logic. Well, it depends on our conception of logic in general. Some people think that the notion of logic should be reserved for systems which come with a fully-fledged and recursive proof theory, and a model-theoretic semantics. On this narrow conception of logic, most nonmonotonic logics do not count as a proper logic. But more liberal conceptions of logic are entertained as well, according to which any inference system constitutes a logic. In essence, an inference relation maps sets of sentences to sets of sentences such that this mapping is guided by some idea of truth preservation: if all members of a given set of premises are true, then all sentences inferable from this set must be true as well. Moreover, there are inference relations which are guided by a weaker requirement: if all members of a given set of premises are believed, then it is rational to believe a certain set of conclusions. Nonmonotonic logics and formal theories of belief revision define inference relations in this sense. What matters for our investigation is that the inference relation $\vdash_a$ is well defined. We will investigate general logical properties of this relation elsewhere.

4 Abductively Redundant Laws

We studied abductive reasoning in order to capture inferences from the spurious cause to the common cause. The motivation for this is to show that correlations between spurious causes and their effects are redundant in the set of generalizations about the respective causal scenario relative to a system of abductive reasoning. We now want to show that this strategy has been successful. Let us first further specify our notions of redundancy and non-redundancy:

Definition 20. Abductive Redundancy
A sentence ϕ is abductively redundant in a set Γ of sentences iff $\phi \in \Gamma$ and $\Gamma \setminus \{\phi\} \vdash_a \phi$.

Definition 21. Abductive Non-redundancy
A sentence ϕ is abductively non-redundant in a set Γ of sentences iff $\phi \in \Gamma$ and $\Gamma \setminus \{\phi\} \nvdash_a \phi$.

These definitions give rise to the notion of deduction which is based on laws which are abductively non-redundant:

Definition 22. $\Gamma \vdash_N \phi$
Let Γ be a set of first-order sentences and ϕ be such a sentence. We say there is a deduction of ϕ from Γ based on non-redundant laws —in symbols $\Gamma \vdash_N \phi$—iff there is a deduction of ϕ from Γ such that any sentence ψ used to justify an inferential step has the following property: if ψ is an implication or a universal sentence, ψ is abductively non-redundant in Γ.

Unless otherwise specified, we henceforth understand the notion of non-redundancy in the sense of abductive non-redundancy. The next step is to show, for various examples, that there is no forward-directed inferential path from the spurious cause to its effect such that only non-redundant laws are used. Moreover, we need to show that there is such a path from the common cause to its effects. Once this has been shown, we are able to further refine our inferential analysis of causation.

5 Common Causes in Conjunctive Scenarios

Let us begin with the example of freezing water, and introduce some symbolic notation to study this scenario:

- W: there is a reservoir of liquid water.
- B: the air temperature drops below zero degrees Celsius.
- T: there is a thermometer.
- F: the water in the reservoir starts to freeze.
- L: the thermometer indicates an air temperature of less than zero degrees Celsius.

The events of the scenario are governed by two laws:

$$W \wedge B \rightarrow F \tag{λ_1}$$

$$B \wedge T \rightarrow L. \tag{λ_2}$$

The first law, roughly, says that water starts freezing if the air temperature drops below zero degrees Celsius. The second law says that a thermometer indicates an air temperature of less than zero degrees Celsius if the air temperature actually falls below zero degrees Celsius. The implications may be read as shorthand notations for universal sentences. Some assumptions remain implicit, as already indicated in Section 1. We assume, for example, that the thermometer measures the temperature of the air to which the water reservoir is exposed to. It's a simple and idealized model of two related processes: freezing of water and a change of the thermometer reading. For simplicity we leave out the underlying physics, which centres on the kinetic theory of heat and the theory of thermodynamic equilibrium. The causal graph of Figure 64 seems a fair representation of our causal verdicts:

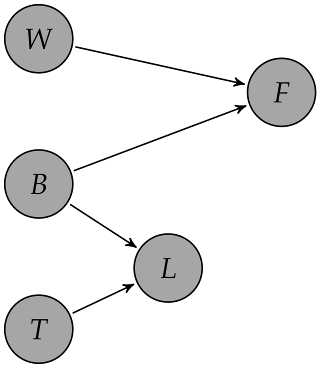


Figure 64: Freezing of water

The drop of temperature is a common cause of two different effects: the freezing of water and the fall of the thermometer reading. It is a common cause, which is embedded in two different conjunctive causal scenarios. Since a thermometer reacts to a fall of temperature before a medium-sized reservoir of water starts to freeze, event L precedes event F. Now, what about the following law?

$$L \wedge W \rightarrow F. \tag{λ_3}$$

If a thermometer indicates an air temperature of less than zero degrees Celsius and a reservoir of liquid water is present, then this water starts to freeze. *Ceteris paribus*, this law holds true of the causal scenario in question inasmuch as the other two laws do. Also, the Humean convention is satisfied. However, we do not think that λ_3 represents a causal relation. A thermometer reading is not a cause for water to freeze.

Our unificationist proposal helps draw the distinction between causal and non-causal laws. We can show that λ_3 is abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. For this to be seen, start a subproof from $L \wedge W$. This gives us L. Since L can be inferred from $B \wedge T$ deductively and in a forward-directed manner, we can infer $B \wedge T$ from L by our inference rule Abduction. $B \wedge T$ gives us B by classical logic. Further, B and $L \wedge W$ imply $W \wedge B$ by classical logic. Using λ_1 we can infer F from $W \wedge B$ by classical logic. This completes the subproof from $L \wedge W$ to F. By Implication Introduction, this subproof lets us infer $L \wedge W \rightarrow F$.

Of course, we also need to show that λ_1 and λ_2 are not abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. Suppose, for contradiction, λ_1 is abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. Then there is a subproof from $W \wedge B$ to F, in which abductive and classical inferences are admitted, and λ_2 and λ_3 are given as premises in the superordinate proof. It is easy to show that F is not a classical logical consequence of the set $\{W \wedge B, \lambda_2, \lambda_3\}$. Hence, F cannot be derived by classical logic in the subproof from $W \wedge B$. Abductive inferences are therefore needed to infer L from $W \wedge B$. To use λ_2 or λ_3 in an abductive inference, we must infer L or F from $\{W \wedge B, \lambda_2, \lambda_3\}$ by classical logic. However, neither L nor F is a classical logical consequence of the set $\{W \wedge B, \lambda_2, \lambda_3\}$. Hence, neither L nor F can be derived from the set $\{W \wedge B, \lambda_2, \lambda_3\}$ by classical logic. Thus we have obtained a contradiction. Analogously, we can show that λ_2 is not abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$.

Note that our result about the freezing of water has some general significance. As observed above, the common cause is embedded in two different conjunctive scenarios. First, the presence of liquid water and temperatures below zero degrees are conjunctive causes for water to freeze. Second, the presence of a thermometer and temperatures below zero degrees are conjunctive causes of temperature measurements below zero degrees. It is furthermore obvious that the above logical demonstration does not depend on the specific meaning of the propositional variables W, B, T, L, and F.

Hence, we have shown that our proposed solution to the problem of spurious causation works for all causal scenarios where a common cause is embedded in two different conjunctive causal scenarios.

At the same time, we have to observe that our proposal fails to work for very simple accounts of common causes. Consider the following account of freezing water:

$$B \to F \qquad (\lambda_1')$$

$$B \to L \qquad (\lambda_2')$$

$$L \to F. \qquad (\lambda_3')$$

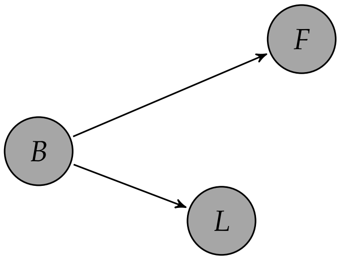


Figure 65: Freezing of water simplified

We take the law $L \to F$ to represent a spurious causal relation. However, each of the implications λ_1', λ_2', and λ_3' is abductively redundant in the set $\{\lambda_1', \lambda_2', \lambda_3'\}$. This poses the question of which representation of the causal scenario we should choose. The implication $W \wedge B \to F$ strikes us as more accurate than the implication $B \to F$. The latter implication is obviously not correct if the temperature is well below zero degrees, but no water is present. Such conditions obtain at night in certain deserts. We should therefore interpret the natural language conditional ‘whenever the temperature is below zero degrees Celsius, then water starts to freeze’ as short for the conditional ‘whenever the temperature is below zero degrees Celsius and there is a reservoir of liquid water, then this reservoir starts to freeze’. We can therefore justify our choice of the conjunctive model over the non-conjunctive one by pointing out that the statements of the former are more accurate than those of the latter.

6 Independence of Causes

Notice that all causes of the above common cause scenario are independent of one another. The presence of a thermometer is independent of the air temperature. Likewise, the presence of a reservoir of liquid water is independent of whether or not there is a thermometer nearby. There is some negative correlation between the air temperature being below freezing point and the presence of a reservoir of liquid water. This correlation, however, is not deterministic at all. For it takes some time until a reservoir of liquid water turns to ice completely. In most geographical regions, lakes never freeze all the way down to the bottom, even during longer periods of very low temperatures.

The last observation is important for the following reason. If said causes were not independent of one another, our approach to spurious causation would fail for the scenario in question and similar ones. Suppose we have a common cause which is embedded in the two conjunctive scenarios, as described in the above section. Further, let us suppose that the causes of each conjunctive scenario have a common cause. Applied to the above scenario of freezing water: W and B have a common cause C_1. B and T have a common cause C_2. Moreover, let us suppose C_1 and C_2 have a common cause C_0. The causes W, B, and T are thus interdependent, and connected by common causes:

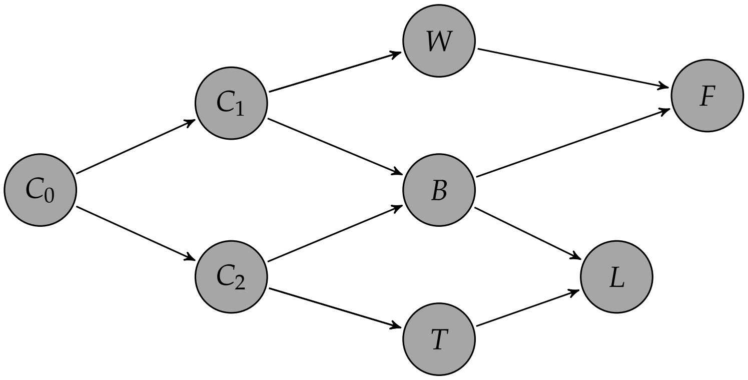


Figure 66: Interdependence of causes in a common cause scenario

The additional causal connections give rise to the following implications:

$$C_0 \rightarrow C_1 \wedge C_2 \tag{λ_4}$$
$$C_2 \rightarrow B \wedge T \tag{λ_5}$$
$$C_1 \rightarrow W \wedge B. \tag{λ_6}$$

It can then be shown that λ_1, λ_2, and λ_3 are abductively redundant in the set $\{\lambda_1, \ldots, \lambda_6\}$. This is a problem since λ_1 and λ_2 are presumed to be genuinely causal regularities, while λ_3 is thought of as a mere spurious one. Put differently, we cannot distinguish between genuine and spurious causes anymore using the criterion of abductive non-redundancy. Of course, this consideration is just hypothetical. As a matter of fact, there are no deterministic correlations among the causes W, B, and T. Hence, our approach to spurious causation works for the present scenario and similar ones.

These considerations have interesting connections to Hausman's (1998) independence theory of causation. On this theory, the core of both probabilistic and deterministic causation is as follows: any effect has at least two causes, and these causes are causally independent in the sense of not having a common cause. Hausman has shown for counterfactual and interventionist approaches to causation that they tacitly rely on this independence principle. We are happy to point out that our theory of causation relies on this principle too—at least in common cause scenarios. If the principle did not hold in such scenarios, our approach to spurious causation would fail to work. Validity of the principle is understood with respect to our concrete causal judgements.

7 Common Causes in Disjunctive Scenarios

Suppose a common cause is part of two disjunctive scenarios. Let us represent such a scenario by simple implication laws at an abstract level:

$$A \vee C \rightarrow E \tag{λ_1}$$
$$C \vee B \rightarrow D. \tag{λ_2}$$

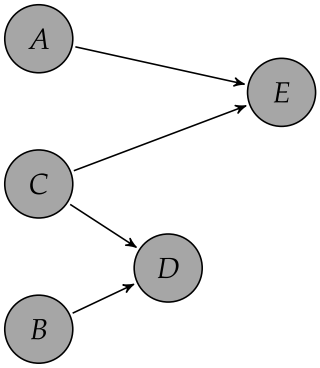


Figure 67: Common cause embedded in two disjunctive scenarios

Further, suppose the antecedent events of these implications precede the respective event of the consequent. By this assumption we can apply the Humean convention, and interpret these laws causally. Let us also assume that there are no events other than C and B which are regularly connected with D. D precedes E. In such a scenario, the following implication holds true as well:

$$D \wedge \neg B \rightarrow E. \qquad (\lambda_3)$$

A causal interpretation of this law seems counterintuitive, even though it satisfies the Humean convention. For example, when we interpret Figure 67 as neuron diagram, we do not think that activation of D and non-activation of B jointly cause the activation of E. Mackie (1980, p. 81-4) interprets a famous example of spurious causation by Russell (1921/2009, p. 289) along the lines of two interrelated disjunctive scenarios. We will discuss this example below, but remain at the abstract level for the logical analysis to follow.

We show that our proposal for the demarcation between spurious and genuine causes works well for the present causal scenario. The spurious causal relation $D \wedge \neg B \rightarrow E$ is abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. For this to be seen, start a subproof from $D \wedge \neg B$. This gives us D by classical logic. By Abduction, we obtain $C \vee B$ from D. $D \wedge \neg B$ gives us $\neg B$ by classical logic. From $B \vee C$ and $\neg B$ we infer C. Using λ_2, we infer E from C. We have thus derived E from the assumption $D \wedge \neg B$. By Implication Introduction, we infer $D \wedge \neg B \rightarrow E$.

The genuine causal relation $A \vee C \rightarrow E$, by contrast, is not abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. If we start a subproof from $A \vee C$, we cannot derive E, even if λ_2 and λ_3 are given as premises and abductive inferences are admitted. Suppose, for contradiction, we can infer E from $A \vee C$ in a subproof. It is easy to show that E is not a classical logical consequence of the set $\{A \vee C, \lambda_2, \lambda_3\}$. Hence, abductive inferences are needed to derive E from $A \vee C$ in the subproof. To use λ_2 or λ_3 for such an inference, E or D must be derived from $\{A \vee C, \lambda_2, \lambda_3\}$ by classical logic. However, it is easy to show that neither E nor D is a classical logical consequence of the set $\{A \vee C, \lambda_2, \lambda_3\}$. Hence, neither E nor D can be derived from the set $\{A \vee C, \lambda_2, \lambda_3\}$ by classical logic. Thus we have obtained a contradiction.

In a manner analogous to this demonstration, we can show that the genuine causal relation $C \vee B \rightarrow D$ is not abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$. Our proposal for demarcating the distinction between genuine and mere spurious causes thus succeeds: there are forward-directed inferential paths from A and C to E—as well as from B and C to D—such that all laws used in the path are non-redundant. But there is no such path from the spurious cause $E \wedge \neg A$ to D. This is as it should be, provided we agree that A, B, and C are genuine causes, while $E \wedge \neg A$ is a mere spurious cause.

It is worth looking at a concrete example of interrelated disjunctive causes. Imagine two factories where the end of the shift is signalled by hooters. Suppose the knocking-off time at the two factories is five o'clock. One factory is in Manchester, the other in London. Then the sounding of factory hooters in Manchester is regularly followed by London workers leaving their work. And yet we do not consider the former event a genuine cause of the latter. This is a famous example of spurious causation due to Russell (1921/2009, p. 289). Mackie (1980, pp. 81–4) interprets the example along the lines of two interrelated disjunctive scenarios. Drawing on this interpretation, we suggest the following formal account:

- F: It is five o'clock according to Greenwich time.
- T_M (T_L) : the hooters at the Manchester (London) factory are tested.
- H_M (H_L): the hooters at the Manchester (London) factory sound.
- L_M (L_L): The Manchester (London) workers leave their work.

The causal scenario is governed by the following laws:

$$T_M \vee F \to H_M \tag{λ_1}$$

$$T_L \vee F \to H_L \tag{λ_2}$$

$$H_M \wedge F \to L_M \tag{λ_3}$$

$$H_L \wedge F \to L_L. \tag{λ_4}$$

On this account of the causal scenario, the event that it is five o'clock and the sounding of hooters at the Manchester factory are causes of the workers to leave their work at the Manchester factory shortly after five o'clock. Likewise for the workers at the London factory. For these causal relations, we have an inferential connection between cause and effect, and the cause precede its effect. However, the following law seems to hold true as well:

$$\neg T_M \wedge H_M \to L_L. \tag{λ_5}$$

In words, whenever the hooters at the Manchester factory sound without being tested, the workers at the London factory leave their work. Again, we can show that there is an asymmetry between the laws $\lambda_1, \ldots, \lambda_4$ on the one hand, and λ_5 on the other. These laws are not on a par with one another. λ_5 is abductively redundant in the set $\{\lambda_1, \ldots, \lambda_5\}$, while none of $\lambda_1, \ldots, \lambda_4$ is abductively redundant in this set. This can be shown in a manner analogous to the above demonstration in this section. Hence, our strategy to demarcate between spurious and genuine causes succeeds again. There is no inferential path from the sounding of hooters at the Manchester factory to the London workers leaving their work such that the path is forward-directed and all laws are non-redundant.

One objection to our formal account of the causal scenario is worth considering. One may well argue that hooters at the Manchester factory are not tested so often, so when they sound, it is almost always five o'clock. Hence,

$$H_M \to L_L \tag{λ_6}$$

is a pretty good *ceteris paribus law*. Other things being equal, the London workers leave their work when the Manchester hooters sound. And this law is not abductively redundant in the set $\{\lambda_1, \ldots, \lambda_4, \lambda_6\}$. We reply to this objection by admitting that *ceteris paribus laws* are perfectly fine and acceptable for modelling deterministic causal relations. At the same time, we are free to prefer the more accurate model in case two different models of a scenario yield different causal verdicts, while one model is more accurate than the other. Accuracy is in this part of our investigation understood with respect to the non-causal claims of the respective models. The sets $\{\lambda_1, \dots, \lambda_4\}$ and $\{\lambda_1, \dots, \lambda_5\}$ have fewer exceptions than the set $\{\lambda_1, \dots, \lambda_4, \lambda_6\}$. Hence, we have reason to prefer the former sets over the latter as an account of the causal scenario.

Notice, furthermore, that our formal description of the scenario may be refined in several ways. For example, it would be more accurate to represent the event that it is five o'clock as a conjunctive factor among other conjunctive factors which together cause the hooters to sound. In addition to the event that it is five o'clock, there needs to be an agreement between workers and the owner of the respective factory that knocking-off time is five o'clock. Also, electrical power needs to be available at five o'clock for the hooters to sound. And so on. Going more fine-grained in this way allows us to even better distinguish genuine from spurious causes.

8 Conjunctive and Disjunctive Scenarios Combined

Finally, we must wonder if our approach to spurious causation works for common causes embedded in a combination of a conjunctive with a disjunctive scenario. We can describe such a scenario at the abstract level as follows:

$$A \wedge C \rightarrow E \tag{λ_1}$$

$$C \vee B \rightarrow D. \tag{λ_2}$$

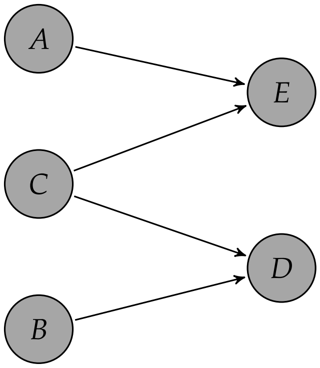


Figure 68: Combination of a disjunctive with a conjunctive scenario

If E precedes D, the following implication stands for a spurious causal relation:

$$E \rightarrow D. \tag{λ_3}$$

If, by contrast, D precedes E, the following implication may be considered a spurious causal relation:

$$A \wedge D \wedge \neg B \rightarrow E. \tag{λ_4}$$

Fortunately, all goes well. We can show that λ_3 is abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3\}$, while λ_1 and λ_2 are not. Likewise, λ_4 is abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_4\}$, while λ_1 and λ_2 are not. Finally, λ_3 and λ_4 are abductively redundant in the set $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$, while λ_1 and λ_2 are not.

We can prove these results using the pattern of the above proofs for common causes embedded in purely conjunctive and purely disjunctive scenarios. Hence, there is no forward-directed inferential path from the spurious cause to its effect such that all laws of this path are not abductively redundant. There is an inferential path from the spurious cause to its effect, which however uses an abductively redundant law. Our analysis thus properly discriminates between genuine and spurious causes in the present scenario.

9 Lightning

Why are we hesitant to call the flash in the sky a cause of the thunder? We think that the flash in an event of lightning and the thunder have a common cause: an electrostatic discharge through the atmosphere between a cloud and the ground. Since light travels faster than sound, we see the flash first, and then we hear the thunder. That is, the spurious cause precedes its effect, and a simple Humean solution to the problem is not available. How does our approach to spurious causation fare with this scenario?

First, we must wonder if there are further causal factors in play that form a conjunctive or disjunctive scenario together with the common cause. If so, the above results about common causes in conjunctive and disjunctive scenarios apply to the electrostatic discharge in lightning. It seems as if there are such factors. To put it more carefully, we can describe an event of lightning such that further causal factors are revealed. The flash in the sky is an optical phenomenon which occurs to an observer only if the sky is visible to her or him. If I am sitting in my apartment with the screens of all the windows closed, I will not see any flash at sky, despite the electrostatic discharge between the cloud and the ground. Or, when I am surrounded by sky scrapers, only a tiny fraction of the sky is visible to me so that a flash may escape my noticing. Similar considerations apply to the thunder. If I am sitting in a room with soundproof walls, such as a studio for producing music, I will not hear the thunder.

One may object to this description that it puts too much emphasis on the direct observation of flash and thunder. We should rather describe the flash in terms of certain electromagnetic waves and the thunder in terms of certain acoustic waves. However, even the propagation of electromagnetic waves depends on certain conditions. Once such waves originate from the area where electricity flows from the cloud to the ground, they will reach a certain spatiotemporal location in the vicinity of the discharge only if this location is not screened off by an opaque object, such as a building or a wall. Likewise, acoustic waves reach such a location only if they have a medium to travel. Sound waves are not able to reach, for example, an area of artificially produced vacuum. We see that the presence of air is a conjunctive factor for the presence of certain sound waves originating from the area of the discharge. And the absence of opaque objects is a conjunctive factor for the presence of certain electromagnetic waves originating from the discharge.

However, we can also describe the effects of lightning in a manner that no conjunctive factors are needed. Whenever there is an electrostatic discharge between a cloud and the ground, then there are always spatiotemporal regions where the electromagnetic waves—emitted by the channel of the electric current—can freely travel. So it holds extensionally universally true that, whenever electromagnetic waves of a specific wave length and intensity are emitted, then certain acoustic waves are emitted from the same source. Even on the ground, we always find spatiotemporal locations where both the flash of a lightning can be seen and the thunder be heard. In this description, the common cause is not embedded in any conjunctive or disjunctive scenario. Our abstract solution to the problem of spurious causation fails to apply then. Let us therefore take a closer look at the physics of lightning, and see if our analysis of spurious causation can be shown to succeed independently of an embedding of the common cause in conjunctive or disjunctive scenarios.

To explain and to describe lightning at a more fundamental and detailed level of physics is far from trivial.

> Despite being one of the most familiar and widely recognized natural phenomena, lightning remains relatively poorly understood. … The study of lightning and related phenomena involves the synthesis of many branches of physics, from atmospheric physics to plasma physics to quantum electrodynamics, and provides a plethora of challenging unsolved problems. (Dwyer and Uman 2014, p. 147)

And yet, we should try to say more about lightning in order to further confirm our account of spurious causation. Many of the unsolved problems concern the conditions under which an electric current can flow to the ground in the first place. They concern details of the development of such a current. Once we assume that electric current—of a certain magnitude—flows from a cloud to the ground, the physics of lightning is relatively well understood. This electric current is the common cause of flash and thunder.[4]

[4] The following description of lightning is mainly based on Rakov and Uman (2003), and Dwyer and Uman (2014), which are also recommended for further reference.

Notably, we cannot see or hear the electric current. Neither the flash nor the thunder is directly caused by the electric current. The flow of electric current rather leads to a sudden and extreme rise of temperature in those regions of the air where electricity is conducted. The channels conducing electricity reach temperatures of around 30,000 Kelvin. Why does the electric current heat up the air? Electrons moving toward the ground collide with ions and molecules of the air. These collisions transmit kinetic energy to the ions and molecules. According to the kinetic theory of heat, kinetic energy amounts to thermal energy so that the temperature rises where electrons flow in a specific direction.

The flash and the thunder are then caused by an extreme rise of temperature in the channels of the air that conduct electricity. What we see at the sky are these channels. Two processes are responsible for the emission of visible light. First, black-body radiation. Any macroscopic physical object emits electromagnetic waves whose frequency depends on its temperature. If the temperature is above a certain threshold, we can see the radiation. To put it more technically, the emitted frequency is within the visible spectrum of electromagnetic waves.

The other process which makes the electricity conducing channels visible are excitations of gas molecules. Kinetic energy of such molecules is transformed to excitations of outer electrons, that is, electrons move to higher energy levels. When these electrons drop back to their initial energy state, photons of light are emitted. The frequencies of light emitted through this process depend on the atomic properties of the respective gas. More generally, chemical elements have specific frequencies of light emission. Since a great deal of the atmosphere consists of nitrogen, the frequencies observed for a flash in the sky are largely due to excitations of nitrogen atoms.

Why do we hear a thunder when an electric current of about 30,000 Ampere flows to the ground? As with the flash, it is the extreme rise of temperature that accounts for the thunder. The kinetic theory of heat tells us that temperature translates to kinetic energy of the molecules and atoms. When the temperature of the electricity conducing channels suddenly rises, the accelerated motion of gas molecules creates a shock wave that we hear as thunder. Put more technically, the accelerated motion leads to a sudden increase of pressure within the channels where electricity flows. This sudden increase of air pressure creates an acoustic wave.

The propagation of acoustic waves is quite well understood, and follows equations analogous to those governing the propagation of electromagnetic waves. Recall that it is electromagnetic waves in terms of which physicists describe optical phenomena visible to us. The propagation of light can be accounted for by the equations of electrodynamics.

Now that we know a bit more about the physics of lightning, let us resume our discussion of spurious causation. We need to show that there is no forward-directed inferential path from the flash in the sky to the thunder such that the laws of this path are not abductively redundant. To make the problem harder and more interesting, let us assume that there are no obstacles to the propagation of acoustic and electromagnetic waves other than the Earth itself.

If we consider some of the physical theories underlying the phenomenon of lightning, it seems impossible to devise a formal proof that the implication between flash and thunder is abductively redundant, while the other laws are not. The available accounts of the physics of lightning are far away from the deductive nomological ideal, according to which we can derive the development of a physical system over time starting from boundary conditions and universal laws as premises (see Rakov and Uman (2003), and Dwyer and Uman (2014)). It is only certain idealized systems, such as a single nitrogen atom, for which our theoretical accounts conform to the formal deductive picture. The accounts of lightning in physics combine theoretical results about such idealized systems with experimental findings about actual lightnings and coarse-grained descriptions of various physical systems, such as the atmosphere.

And yet, when studying the accounts of lightning in physics, we can recognize inferential paths among the various events involved. At the core of lightning is the event of an electrostatic discharge through the atmosphere between a cloud and the ground. Hence, there is an electric current flowing from the cloud to the ground. The electrons making up this current collide with the molecules of the air so that these molecules gain kinetic and thermal energy. By our theories of ideal and real gases, the rise of temperature leads to increased pressure. Also, atomic theory tells us that kinetic energy of atoms and molecules gets transformed into excitations of electrons. Furthermore, we know by the laws of atomic theory that electromagnetic radiation is emitted when excited electrons drop back to their initial energy state. And so on.

We can thus recognize a sequence of events such that each element in the sequence results from its predecessor (if there is one), and the inferential connection between the two events is backed up by laws of well established physical theories. No element in the sequence is temporally later than its successor. At least some elements are temporally later than their predecessor. Hence, we can say that there is a forward-directed inferential path from the electrostatic discharge to the emission of bright light, and one from the discharge to the shock waves creating thunder. These paths make essential use of laws from well established physical theories, including most prominently electrodynamics, acoustics, atomic physics, and thermodynamics. They also use general descriptions of the macroscopic physical systems involved, in particular the Earth's atmosphere.

Now, our central claim is that none of the laws used to infer the bright flash in the sky from the discharge becomes abductively redundant if we add to the set of these laws the implication that thunder follows the flash in the sky. There is no way, for example, to infer any law of atomic theory from the implication between flash and thunder, even if other laws of well established physical theories remain available as premises. Similar considerations apply to the inferential path from the discharge to thunder. By contrast, the implication between flash and thunder can be inferred by abductive and deductive reasoning. For this to be seen, suppose we see a bright flash in the sky. In view of our physical theories and our knowledge in geophysics, there is only one type of forward-directed inferential path that concludes with the event of such an observation. This path starts with the assumption of an electrostatic discharge through the atmosphere between a cloud and the ground. There is no other type of event from which we could infer the observation of a bright flash in the sky, set aside events that precede or proceed the occurrence of such a flash. Hence, by Abduction, there occurred an electrostatic discharge between a cloud to the ground. By Implication Introduction of classical logic and the deductive inference from the electrostatic discharge to the propagation of an acoustic shock wave, we can infer the implication that, when a bright flash occurs at the sky, thunder will occur as well.

10 Trivial Spurious Causes

Day regularly follows night, but night is not considered a cause of day (Mill 1843/2011, Book III, § 6). Low tides are followed by high tides, without the former being seen as a cause of the latter. Life is followed by death, and yet we do not think that life causes death. In all of these examples, we can infer an event from another in a forward-directed manner, and the inferred event occurs later than the event from which the inference started. But it seems wrong to say that there is a causal connection between the two events. What is wrong?

We must wonder whether there are common causes in the background. If so, we could try to subsume the present examples under the above analysis of spurious causation. However, the structure of inferential paths seems to differ from that of the common cause scenarios in the previous section. We know that the cycle of night and day is caused by rotation of the Earth and emission of light by the Sun. But there is a difference between causing the cycle of night and day, and causing a specific event of it's being day at a certain time and place. We cannot infer the event of it's being day at a specific time and place from the statement that the Earth is rotating and the Sun is emitting light. Nor can we infer an event of high tides from the fact that the Moon is orbiting around the Earth.

Suppose there is daylight at a specific time t and place p on the Earth. We then have two competing causal explanations. One says that the daylight is caused by the preceding night. The other says that the daylight is caused by the Sun emitting light, and there being no light-absorbing obstacles between the Sun and place p at time t other than the atmosphere. Obviously, the latter explanation in terms of sunlight is by far more convincing and appealing. Why so?

Again, unificationist ideas help us clarify things. We have a theory about our planetary system which allows us to explain a number of different phenomena: the cycle of day and night, the daily cycle of different angles of sun rays, the cycle of seasons, polar night and polar day, the presence of Coriolis forces on the surface of the Earth, eclipses of the Moon and the Sun, and so on. This theory says that the Sun is at rest, and planets are orbiting around the Sun. Most notably, the Earth itself is rotating, and the angle between the axis of rotation and the orbit of the Earth around the Sun is around 23.5

degrees. The Sun emits electromagnetic radiation in the visible spectrum. This theory of our planetary system comes in different degrees of precision. We have different formulations, which vary with respect to their degree of detail and quantitative information. For the discussion to follow it does not matter if we view the theory as a single theory or a family of theories.

Clearly, the causal explanation of daylight in terms of sunlight proceeds from premises which have a lot of unificatory power. As just emphasized, we can infer from the theory of our planetary system a number of different phenomena. By contrast, very little can be inferred from the generalization that day follows night. Most notably, this generalization can be inferred from the theory about our planetary system. It is therefore deductively redundant relative to some syntactic formulation of this theory. Suppose T is some syntactic formulation of the theory of our planetary system. Let λ be the law, or generalization, that day follows night. Then λ is deductively redundant in the set $T \cup \{\lambda\}$.

It seems as if our approach to spurious causation continues to work for the present example: C is a mere spurious cause of E iff there is a forward-directed inferential path from C to E, C precedes E, but the inferential path contains some redundant law. Put differently, there is no forward-directed inferential path from the spurious cause to its effect such that all laws of this path are non-redundant. Only genuine causes have forward-directed inferential paths to their effects such that all laws of the path are non-redundant. The inference from night to daylight is forward-directed, but rests on a redundant generalization. Hence, night is a spurious cause of daylight.

However, there remains to consider a subtle objection to our analysis of trivial spurious causes. Certainly, we cannot infer—by deductive classical reasoning—any element of the theory of our planetary system from the generalization that day follows night. Nor can we infer abductively any such element from the latter generalization alone. However, it seems as if we can abductively infer that the Earth is rotating from the generalization that day follows night and certain other parts of the theory of our planetary system. Suppose we know that it is day at a specific place p whenever this place is exposed to sunlight. By contrast, it is night at p whenever p is in the shadow of the Earth with respect to sunlight. Further, we know or assume that the Sun is at rest. Also, we know that it takes the Earth a year to do a complete orbit around the Sun. This information excludes that the cycle of night and day can be inferred from the motion of the Earth around the

Sun, given that the length of night and day is less than 24 hours (at least in geographical regions between the polar circles). Then there is a forward-directed inferential path from the premise that the Earth is rotating to the observation that day follows night. Given our background knowledge, the rotation of the Earth is the only explanation of the cycle of night and day. By Abduction, we infer that the Earth is rotating around its own axis. Hence, there is a member of T (the theory of our planetary system) which becomes abductively redundant if we add λ (the law that day follows night) to T.

Why is this result a problem? The rotation of the Earth is a genuine cause of a number of different phenomena, among which the cycle of night and day is just one. Even worse, the rotation of the Earth is used in genuine causal explanations. A case in point is the dynamics of low-pressure area systems, for which Coriolis forces are an important causal factor. Since we think that causal inferential paths use only non-redundant laws, we do not want to view the rotation of the Earth as a redundant law.

It may be tempting to solve the problem by arguing that the rotation of the Earth is a fact rather than a law. After all, we can represent the claim in question by a simple atomic sentence which says that the Earth is rotating. But it is more correct to spell out the claim in question by a sentence which says that, for any time t, the angular velocity of the Earth has a certain constant value. Then the first-order translation of the latter formulation has the form of a universal sentence. It therefore qualifies as law in the sense of our minimalist and syntactic understanding of laws. Hence, we better do not defend our explication of spurious causation by arguing that the rotation of the Earth is not a law.

Here is a more promising strategy. After all, the generalization that day follows night and the claim that the Earth is rotating are not on a par as regards their inferential power. While the former is deductively redundant, the latter is only abductively redundant—in the context of the theory of our planetary system. Suppose ρ is the claim that the Earth is rotating with a constant angular velocity. T is the theory of our planetary system. λ is the generalization that day follows night. Recall that $\rho \in T$, while $\lambda \notin T$. Then ρ is abductively redundant in the set $T \cup \{\lambda\}$. By contrast, λ is deductively redundant in this set. This leads to the following observation. In the present set of causal scenarios, the (forward-directed) inferential path between the spurious cause and its effect uses deductively redundant laws. The (forward-directed) inferential path between the genuine cause and its effect also uses redundant laws, but these laws are only abductively redundant. By contrast, in common cause scenarios, only the forward-directed inferential path between the spurious cause and its effect uses redundant laws, and these laws are abductively redundant. The inferential paths between genuine causes and their effects are free of any redundant laws. In the next section, we will generalize our approach to spurious causation such that both types of causal scenarios are captured.

11 An Ordering of Unification

In the previous sections, we have studied a number of different scenarios of spurious causation. There are subtle differences between these scenarios, and so it proved difficult to capture all of them in a unified account. Nonetheless, our results give rise to a general observation about the distinction between spurious and genuine causes: the forward-directed inferential paths between genuine causes and their effects are based on laws which are—in some sense—more unificatory than the laws used in the forward-directed inferential paths between spurious causes and their effects. Unificatory power is understood in terms of inferential power. If a law is redundant in the context of a theory, it has less inferential power than the non-redundant members of this theory.

This inferential diagnosis of spurious causation worked particularly well for scenarios with common causes. But even in the other cases, there seems to be a difference in inferential power between the laws used to infer effects from genuine causes compared to the laws used in the inferential paths from the spurious cause. As emphasized above, we can infer from the rotation of the Earth and suitable background theories a number of different phenomena: the cycle of day and night, the daily cycle of different angles of sun rays, the cycle of seasons, polar night and polar day, the presence of Coriolis forces on the surface of the Earth. Such forces, in turn, are causal factors in the dynamics of low-pressure systems and other meteorological phenomena. These inferential explanations are usually presented in a deductive fashion.

By contrast, if we start from the generalization that day follows night, we cannot infer, for example, the presence of Coriolis forces by deductive reasoning alone. To infer such forces, we would have to first infer the rotation of the Earth from the generalization that day follows night by abductive reasoning. Then we could go on to infer the presence of Coriolis forces on the surface of the Earth by deductive reasoning. However, this combination of abductive and deductive inferences is more complex than the deduction of the law that day follows night from the rotation of the Earth. Such combinations are not considered explanatory. We do therefore not say that the cycle of night and day inferentially explains the presence of Coriolis forces on the surface of the Earth. The rotation of the Earth has more explanatory inferential power than the claim that day follows night because the former has simpler inferential connections to a variety of different phenomena.

In light of this observation, we suggest distinguishing degrees of redundancy, depending on the logical means needed to show that a given sentence is redundant. Suppose we have a set L of laws such that $\lambda_1, \lambda_2 \in L$. λ_1 is deductively redundant in L, while λ_2 is abductively redundant. We then say that λ_1 is redundant in L to a higher degree than λ_2 is. This proposal is motivated by the fact that λ_1 can be recovered from $L \setminus \{\lambda_1\}$ by simpler types of inferences than λ_2 from $L \setminus \{\lambda_2\}$. For deductive reasoning is simpler than abductive reasoning, which is obvious from our account of abductive reasoning in Section 3. We can therefore say that $L \setminus \{\lambda_1\}$ is a simpler representation of L than $L \setminus \{\lambda_2\}$ is. When forced to choose between $L \setminus \{\lambda_1\}$ and $L \setminus \{\lambda_2\}$, it is reasonable to go for the former set.

Another way of motivating degrees of redundancy is to emphasize that abductively redundant laws are more central to the inferential power of a theory than deductively redundant ones. In our schematic example, $L \setminus \{\lambda_1\}$ and $L \setminus \{\lambda_2\}$ are on a par with respect to abductive reasoning, but $L \setminus \{\lambda_1\}$ is deductively inferentially more powerful than $L \setminus \{\lambda_2\}$. Hence, λ_1 is easier to dispense with than λ_2. We therefore say that λ_1 is redundant to a higher degree in L than λ_2 is. In brief, deductive redundancy is stronger than abductive redundancy. The claim that day follows night is redundant to a higher degree than the rotation of the Earth—in the theory of our planetary system joined with the claim that day follows night.

Let us now merge the above analysis of spurious causation in common cause scenarios with the present analysis of trivial causal relations. For this to be achieved, we define an ordering of unification among sets of sentences which inferentially represent a given set L of laws. $Cn(\Gamma)$ designates the classical inferential closure of Γ, as is standard. $Cn_a(\Gamma)$, by contrast, the abductive inferential closure of Γ.

Definition 23. $L' >_u L''$ (preliminary)
Suppose L' and L'' are subsets of a set L of laws. L' is a more unified account of L than L''—in symbols: $L' >_u L''$—iff

(1) $|L'| < |L''|$, while $L \subseteq Cn_a(L')$ and $L \subseteq Cn_a(L'')$, or

(2) $|L'| = |L''|$,while $L \subseteq Cn(L')$, but $L \not\subseteq Cn(L'')$.

This definition says that there are two ways in which a subset of a given set L of laws may be more unified than another. First, L' may contain fewer members than L'', while all sentences of L are inferable from both L' and L'', respectively. Second, while L' and L'' have the same cardinality, L is contained in the deductive inferential closure of L', but there are sentences in L which cannot be inferred by deduction from L''. Then L' is a simpler account of L, even if L is contained in the abductive closure of both L' and L''. Condition (1) helps us distinguish genuine from spurious causes in common cause scenarios. It also works for the lightning scenario. Condition (2) allows us to draw the distinction in scenarios with trivial causal relations. Note that $>_u$ is a binary relation defined for the set of subsets of a given set of laws.

Suppose different types of redundancy are in play when considering a given set L of laws. That is, L contains redundant laws in the sense of condition (1) and ones which are redundant in the sense of condition (2). Then the above definition cannot be applied directly. However, we can capture different types of unification by considering partitions of a given set L of laws: L' is a more unified account of L than L'' iff there is a partition of L such that it holds for any member of this partition that the intersection with L' is more unified—in the sense of Definition 23—than the intersection with L'', while there is no partition for which the reverse claim holds. For simplicity, we leave partitions implicit in what follows.

The alert reader will have noticed another problem with our proposal for defining an order of unification: we implicitly assume that different explicit beliefs are not lumped together by conjunction. To give a simple example, suppose L_1 contains n different implications, where $n > 1$. Further suppose that $\lambda \in L_1$ is deductively redundant in L_1. Now, let L_2 be the singleton which contains the conjunction of the implications in L_1. Obviously, our definition of $>_u$ gives the intended result for L_1, but not for L_2.

This problem may be addressed by comparing the syntactic size of different sets of sentences instead of their cardinality.[5] The syntactic size of a given sentence may be determined by the cardinality of the set of atomic subformulas. The syntactic size of a set of sentences is then simply given by the arithmetic sum of the syntactic size of each member. In more formal terms, let $A(\phi)$ be the set of atomic subformulas of a given sentence ϕ. Then we define the syntactic size $S(L)$ of a set L of sentences as follows:

$$S(L) = \sum_{\phi \in L} |A(\phi)|.$$

Comparing sets of sentences with respect to their syntactic size results in the following order of unification:

Definition 24. $L' >_u L''$
Let L be a set of laws. Let L' be a set of laws such that any atomic subformula of a sentence in L' is also an atomic subformula of a sentence in L. Likewise for L''. L' is a more unified account of L than L''—in symbols: $L' >_u L''$—iff

(1) $S(L') < S(L'')$, while $L \subseteq Cn_a(L')$ and $L \subseteq Cn_a(L'')$, or

(2) $S(L') = S(L'')$, while $L \subseteq Cn(L')$, but $L \not\subseteq Cn(L'')$.

This is our final analysis of unification, which will be adopted to define the notions of redundancy and non-redundancy. Notice that $>_u$ is a strict partial order on the set of laws which are composed of the atomic subformulas of the sentences in L.[6]

Given an ordering of unification thus defined, we can say what it is for a set of laws to be maximally unified: the maximum of the ordering $>_u$ defines the set of maximally unified accounts of L. That is, L' is a maximally unified account of L iff there is no L'' such that $L'' >_u L'$. For simple causal scenarios, the maximum of $>_u$ is often a singleton. But the maximum of an ordering may contain more than one member. The maximum of $>_u$ is always well defined since we assume that the set of explicit beliefs of an epistemic state is finite.

[5] Friedman (1974) was well aware of this problem. Consequently, he measured unification not in terms of the cardinality of sets of sentences. The suggested alternative measure is based on the notion of a set of independently acceptable sentences.

[6] Recall that a strict partial order is a binary relation which is transitive and irreflexive.

Comparing the number of subformulas of different sets of laws—rather than the cardinality of the subsets of a given set L of laws—adds further complexity. Fortunately, the problem concerning conjunctions of laws hardly ever arises in practice for sets of explicit beliefs. It seems as if our explicit beliefs are governed by some principle of independent acceptability. That is, it holds for most of our explicit beliefs that they are not logically equivalent to a set of sentences such that each member of the set can be accepted independently of any other member.

Notice that both the preliminary and the final explication of unification solve another problem concerning the redundancy of laws. Suppose we are in a scenario with a common cause which is embedded in two conjunctive scenarios. We think the following two laws are non-redundant:

$$A \wedge C \rightarrow E \tag{λ_1}$$

$$C \wedge B \rightarrow F. \tag{λ_2}$$

Obviously, λ_1 and λ_2 are deductively non-redundant in the set $\{\lambda_1, \lambda_2\}$. Now, consider the following implication:

$$C \wedge A \rightarrow E. \tag{λ_3}$$

One could argue that this sentence represents an explicit belief about the causal scenario in question inasmuch as λ_1 does. The problem, then, is that λ_1 and λ_3 are deductively redundant in the set $L_1 = \{\lambda_1, \lambda_2, \lambda_3\}$. So, C would not be a cause of E anymore since there is no forward-directed inferential path from C to E such that all laws of the path are non-redundant. However, λ_1 and λ_3 are a member of some maximally unified account of L_1, even though there is no such account that contains both λ_1 and λ_3. Hence, there is a forward-directed inferential path between C and E such that all laws of the path are in one and the same maximally unificatory account of L_1. By contrast, the spurious causal law $A \wedge F \rightarrow E$ is not a member of any maximally unified account of $L_1 \cup \{A \wedge F \rightarrow E\}$. In this sense, it is a redundant law.

12 Unification and Causation

It is time to put everything together, and to refine our Humean analysis of causation so that it excludes spurious causes. Our inferential account of spurious causation may be summarized as follows: suppose C and E are occurring events, and C precedes E. Then C is a spurious cause of E iff there is a forward-directed inferential path from C to E, but any forward-directed inferential path from C to E makes use of some redundant law. By contrast, C is a genuine cause of E iff there is a forward-directed inferential path from C to E such that any law of this path is non-redundant.

To explain the notion of redundancy of a law, we have defined the notion of a maximally unified account of a set L of laws: L' is a maximally unified account of a set L of laws iff there is no L'' such that L'' is a more unified account of L than L'. This notion of maximal unification is based on an ordering $>_u$ of unification, defined by Definition 24. Redundancy and non-redundancy of a law can now be understood in terms of the membership in a maximally unified account of laws.

Definition 25. Redundancy and Non-Redundancy
Let L be a set of laws explicitly believed in some epistemic state S. A law $\lambda \in L$ is redundant in L iff λ is not a member of any maximally unified account of L. By contrast, $\lambda \in L$ is non-redundant in L iff it is a member in at least one maximally unified account of L.

Let us now impose the requirement of non-redundancy on the inferential path between cause and effect so as to exclude spurious causes. First, we strengthen the conditional $\gg_F$ as follows:

$$A \gg_{FN} C \in K(S) \text{ iff there is } \bigcup(S \div B(A) \vee B(C)), A \vdash_{FN} C. \quad (SRT_{FN})$$

$\vdash_{FN}$ has its obvious meaning: $\Gamma \vdash_{FN} \phi$ says that there is an inferential path which is forward-directed and satisfies the condition that all laws of this path are non-redundant in Γ. $\bigcup(S \div B(A) \vee B(C))$ is the set of explicit beliefs of the epistemic state S after suspension of judgement on A and C.

Second, we require that $C \gg_{FN} E$ must hold for genuine causes C and their effects E. This leads to the following analysis of causation:

Definition 26. Cause
Let C and E be events. C is a cause of E—relative to an epistemic state S—iff

(C1) $C, E \in K(S)$,

(C2) $C \gg_{FN} E \in K_{>}(S)$, and

(C3) C precedes E.

This analysis of causation obviously draws on unificationist ideas about explanation. In essence, the unificationist account says that 'to explain is to fit the phenomena into a unified picture insofar as we can' (Kitcher 1989, p. 500). Both Friedman (1974) and Kitcher (1989) describe unification in terms of derivations among sentences. They thus adhere to a broadly syntactic understanding of scientific theories. Kitcher (1989, p. 500) thinks that unificationist explanations reveal the causal structure of the world eventually. As indicated in the Introduction, there is substantial overlap and convergence between ideas in Kitcher's account of unification and our approach to causation. However, there are also noteworthy differences.

Our proposed theory takes a broadly logical notion of reason to be more fundamental than the notion of causation. Certain reason relations have a causal meaning, others do not. We try to draw the distinction between causal and non-causal reasons in a unified way. The resulting analysis is to *save the phenomena*—given by our concrete causal judgements—to a maximally possible extent. While we are not committed to a full blown unificationist account of explanation, our analysis of spurious causation is motivated by unificationist ideas: the reason relations of genuine causal relations are superior to those of spurious causal relations because the former employ more fundamental laws, that is, laws which have more explanatory power in the unificationist hierarchy. Put differently, we merely say that unification is one dimension according to which a given explanation may be assessed. In the case of spurious causal explanations, this dimension is decisive for discounting them as genuinely causal.

Since we merely exploit some unificationist ideas and notions for our analysis of causations, standard criticisms directed at the unificationist accounts by Friedman (1974) and Kitcher (1989) do not apply to our analysis. Most importantly, we do not say that explanation always has a unificatory structure. Nor do we say that unification always provides us with an explanation. Presumed counterexamples to the unificationist account, therefore, do not affect our approach to spurious causation. A more general concern arises from what has been termed the *the-winner-takes-it-all problem*, though. At least in physics, scientists have come up with more and more unified accounts. As a consequence of this, we would have to consider only the most fundamental theories as properly explanatory. This does not accord with scientific praxis, where explanations are also provided in terms of less fundamental theories. A case in point are explanations in terms of the ideal gas law.[7]

The *winner-takes-it-all problem* does not arise for our analysis of causation. Consideration of more fundamental theories does not change the verdicts about spurious and genuine causes in such a manner that a presumably genuine cause becomes spurious. Take the simple theory about freezing water from Section 5. Suppose we extend this theory by the kinetic theory of heat—covering not only gases, but also liquids and solid states—and other elements of thermodynamics. Then we can translate the event of a drop of the air temperature below freezing point into a corresponding drop of the mean kinetic energy of the molecules in the air. Furthermore, we have an inferential path from the drop of the mean kinetic energy of the air molecules to the formation of a thermodynamic equilibrium between the water reservoir and the air of the environment. This inferential path represents the transfer of kinetic energy from water to air. As a consequence of this energy transfer, the mean kinetic energy of water molecules drops below a certain threshold. Water, therefore, starts freezing. In brief, we obtain a new inferential path from the genuine cause to its effect which is forward-directed, and satisfies the condition that all laws of the path are non-redundant relative to the set of laws of the epistemic state.

The spurious cause, by contrast, remains without such an inferential path. The kinetic theory of heat does not allow us to infer the freezing of water from a thermometer reading below zero degrees Celsius such that the corresponding inferential path is forward-directed and satisfies the condition that all laws of the path are non-redundant. This can be seen from the following considerations. When we adopt the kinetic theory of heat and other elements of thermodynamics, the causal implications λ_1 and λ_2 become redundant. Since the spurious causal implication λ_3 can be inferred by abductive reasoning from $\{\lambda_1, \lambda_2\}$, this implication remains redundant. (λ_1, λ_2, and λ_3 are understood as in Section 5.)

Even though λ_1 becomes redundant, the event of the air temperature dropping below freezing point remains a genuine cause on our analysis. For we can describe this event in terms of a drop of the mean kinetic energy of the

[7] For a detailed criticism of the unificationist account of explanation, including *the-winner-takes-it-all problem*, see Woodward (2003, sects. 8.6–8.10).

air molecules without using any backward-directed inferences. By contrast, the drop of the thermometer reading cannot be translated into a drop of the mean kinetic energy of the air molecules. If we were to infer a drop of the mean kinetic energy of the air molecules from the fall of the thermometer reading, this inference goes backward in time. For no thermometer immediately displays a drop in mean kinetic energy of the surrounding air. This is well known for liquid thermometers, such as old-fashioned mercury thermometers, but holds also true for electric thermometers.

Similar considerations apply to the famous scenario about the hooters of two factories, one in Manchester and another in London. We have discussed this scenario in Section 7. The verdicts of our analysis do not change if we consider the laws of electrodynamics, acoustics, and some pieces of neurophysiology to better understand why the factory workers leave the factory upon a signal by the hooters. Note, finally, that consideration of more fundamental theories in physics enabled us in the first place to formally distinguish between spurious and genuine causes in the scenario of lightning. In sum, considering more fundamental levels improves or does not change the delineation between spurious and genuine causes on our analysis.

One word on the best system account of laws of nature is in order. This account is a close relative to the unificationist approach to explanation. In essence, it says that 'the laws of nature are the true generalizations that best systematize our scientific knowledge' (Cohen and Callender 2009, p. 2). This understanding of lawhood goes back to Mill (1843/2011), Ramsey (1931a, p. 242), and Lewis (1973b, p. 73). Our theory of causation is obviously sympathetic to the best system account of laws, but we are not committed to that account. The present approach to spurious causation begins with a minimalist and syntactic notion of law, which is weaker than the standard notion of a proper law of nature. Our final notion of a nonredundant law may nonetheless be an interesting candidate to explicate the best system account. We leave this for future research.

13 Other Solutions

Let us now briefly compare our proposal for delineating between genuine and spurious causes to other accounts of this distinction. In doing so, we will also look at probabilistic analyses of causation. An important result of this comparison is that principles of parsimony and simplicity are quite frequently used in contemporary solutions to the problem of spurious causation. Our analysis is thus in good company, but does not coincide with any extant account of deterministic causation.

Reichenbach (1956) was the first to give a thorough treatment of common causes in a probabilistic framework. His strategy for solving the problem of spurious causation—in common cause scenarios—may be described as follows. He begins with a proto-theory of probabilistic causation, according to which causation is simply probability raising. In brief, C causes E iff C raises the probability of E. A probabilistic cause makes the occurrence of its effect more probable. For this proto-theory the problem of spurious causation arises. For example, a significant drop of the barometer raises the probability of stormy weather.

The next step is to refine the proto-theory so as to exclude spurious causes. In essence, C is a genuine probabilistic cause of E iff (i) C raises the probability of E, and (ii) there is no event C' such that C' *screens off* C from E, while C' is earlier or simultaneous to C. C' is said to screen off C from E iff C and E become probabilistic independent once we assume that C' occurs. In more technical terms, C' *screens off* C from E iff $P(E \wedge C \mid C') = P(E \mid C') \cdot P(C \mid C')$. The event of a low-pressure system approaching screens off the drop of the barometer from stormy weather, and precedes the latter events. Hence, the drop of the barometer does not qualify as a genuine cause.[8]

Two points seem noteworthy when looking at Reichenbach's solution to the problem of spurious causation from the perspective of our analysis. First, Reichenbach's definition of probabilistic causation makes use of temporal relations among events, just as our inferential analysis of deterministic causation does. The direction of time, in turn, is defined by Reichenbach in terms of statistical and probabilistic concepts without reference to causal notions. Our inferential analysis of deterministic causation is consistent with this definition.

Second, Reichenbach's account of probabilistic causation cannot directly

[8] See Reichenbach (1956, chs. 18–23) for details of his analysis of probabilistic causation. The final definition of probabilistic causation—also referred to as *causal relevance*— is given by Definition 2 on page 204. We have simplified this definition by the assumption that a common cause is always given by a single event, thereby excluding conjunctions of events.

be extended to deterministic causal scenarios. Suppose for contradiction that deterministic causation is just a limiting case of probabilistic causation. Further, suppose A is a deterministic cause of B, which in turn causes E deterministically. And A precedes B, which in turn precedes E. A, B, and E thus form a simple causal chain:

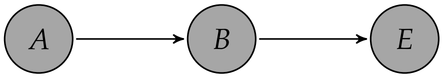


Figure 63: Causal chain

Let us also assume that the unconditional probabilities of the events A, B, and E are between zero and one, which is a natural assumption to make. Then the conditional probabilities $P(B \wedge E \mid A)$, $P(B \mid A)$, and $P(E \mid A)$ are all equal to one. Hence, A screens off B from E. Since A precedes B, this implies that B cannot be a probabilistic cause of E according to Reichenbach's definition of probabilistic causation. Since we assumed that deterministic causation is a limiting case of probabilistic causation, we can infer from this result that B is not a deterministic cause of E. This contradicts our assumption that B is such a cause.

One may defend a generalization of Reichenbach's approach to spurious causation to deterministic scenarios by arguing that simple causal chains do not exist in nature. We may well be able to always find another causal factor of a given effect. This is an implication of Hausman's independence principle, explained in Section 6. We have assumed the latter principles for the effects of a common cause. Our main reason for not using Reichenbach's approach to spurious causation is that the latter uses a different framework. The goal of this book is to analyse deterministic causation, broadly construed.

Unlike Reichenbach (1956), the theory of probabilistic causation in terms of causal models by Spirtes et al. (1993) and Pearl (2000) does not rely on temporal relations among events. While taking the notion of causation more seriously than the mainstream literature in statistics, Spirtes et al. (1993) and Pearl (2000) acknowledge that we cannot derive causal models from statistical data in a unique way. Considerations of simplicity are therefore needed in order to favour one causal model over another with respect to a given set of statistical data. This strategy is in line with the unificationist account of explanation. Both the unificationist account and the best system account of laws of nature are obviously driven by some principle of simplicity. The question thus arises whether modern probabilistic theories of causation could be rephrased so as to fit the unificationist picture of explanation. We leave this question for future research.[9]

Let us now look at proposals of how to solve the problem of spurious causation in a non-probabilistic framework. The analyses of causation by Baumgartner (2013) and Baumgartner and Falk (2019), and Spohn (2006, 2012) merit consideration. There is much convergence between Baumgartner's solution to the problem of spurious causation and our proposal for solving this problem. Both solutions substantially rely on some notion of redundancy, defined for theories about a given causal scenario. Baumgartner's analysis of deterministic causation is a refinement of the INUS account by Mackie (1965). The latter account represents causal relations by biconditionals of the following form:

$$(C_{1,1} \wedge ... \wedge C_{1,n}) \vee ... \vee (C_{k,1} \wedge ... \wedge C_{k,m}) \leftrightarrow E.$$

Each atom $C_{i,j}$ stands for a specific type of event or fact. And each $C_{i,j}$ is an insufficient but non-redundant part of an unnecessary but sufficient condition for E. In this sense, each $C_{i,j}$ is a cause of E at the type level. A type-level cause is at least an INUS condition. As is well known, INUS is an acronym which stands for *insufficient but non-redundant part of an unnecessary but sufficient condition*.

Suppose T is a set of biconditionals of the above form. All members of T are extensionally true. Then each member of T and T itself must satisfy certain conditions of minimality and non-redundancy in order to qualify as a theory of genuine INUS conditions which excludes spurious INUS conditions. In essence, no biconditional must contain a conjunction of factors $C_{i,1} \wedge \ldots \wedge C_{i,l}$ such that the biconditional obtained by removing this conjunction is extensionally true as well. This analysis succeeds in excluding spurious causes in scenarios where each effect of a common cause has at least two causal factors. The two causal factors, one of which is the common cause, may form a conjunctive or a disjunctive causal scenario. We

[9] Woodward's (2003) interventionist account makes use of the framework of causal models by Pearl (2000). The problem of spurious causation is explicitly addressed, and serves as an important motivation for the proposed account. We will discuss this account in Chapter 10.

have discussed such scenarios in sections 5 to 8, and shown that our analysis succeeds in these scenarios as well. We conjecture that theories T which satisfy the minimality constraints in Baumgartner and Falk (2019) are maximally unified in the sense of our unification order $<_u$, defined in Section 11. We leave the proof of this conjecture for future work.[10]

A distinctive merit of the INUS approach to spurious causation is simplicity. It is certainly simpler than our inferential approach. However, the additional complexity of our proposal comes with additional benefits at the level of the overall theory of causation. First, our analysis of causation can be applied to mathematically complex theories in science in a relatively direct manner. What matters are inferences and temporal relations among events which are inferred in an application of the scientific theory to a concrete scenario. The INUS account, by contrast, runs into problems when it is applied to quantitative causal relations. Second, our theory continues to work for causal relations in time-symmetric theories. By contrast, the refined INUS account of causation by Baumgartner and Falk (2019) fails to work for such relations. This will be shown in Section 9 of Chapter 10.

Why is there a problem with quantitative causal relations for the INUS account? Take Newton's law of gravitation. This law tells us that the gravitational force between two objects is proportional to the product of the masses of the two objects, but inversely proportional to the square of their distance. We standardly interpret the law in a causal manner: it's the masses of the two objects which cause the gravitational force between the two. Now, the problem is that there is an infinite set of combinations of masses and distances which determine one and the same gravitational force. To give a simple example, two bodies with the units of mass 1 and 4 exert the same gravitational force upon one another as two objects with the units 2 and 2, provided the distance between the bodies is the same.

If we want to capture the causal relations represented by the law of gravitation in the format of the INUS account, we end up with a set of very complex biconditionals. Each biconditional says that the gravitational force has value F iff the mass of one object has value m_1, that of the other the value m_2, while the distance between the two has value r, or … . For each biconditional, F is a specific real number, and so are m_1, m_2, and r in each

[10] May and Graßhoff (2001) were the first to suggest this refinement of the INUS account of causation. We refer to Baumgartner (2013), and Baumgartner and Falk (2019) because the latter articles go beyond May and Graßhoff (2001) in some respects.

conjunction of the biconditional. Strictly speaking, the biconditionals are not only very complex, but uncountably infinitely long. This is so because we need to use real numbers in classical mechanics. And for each specific value of the force function, there is an uncountable set of combinations of masses and distances of the two objects such that the gravitational force between the two objects has the value in question. Classical first-order logic is certainly not expressive enough for such biconditionals since any classical first-order formula is finite in length. It remains to investigate whether infinitary logics could help. The challenge is that the members of an uncountable set cannot be represented by an infinite sequence.[11]

Newton's law of gravitation cannot directly be adopted into the notation of the INUS account since it does not have the form of a biconditional. A merely qualitative formulation of Newton's law of gravitation would give us only a severely impoverished account of the causal claims in classical mechanics. These considerations generalize to all equations in scientific theories which rely on real numbers. We will look at the law of classical gravitation from the perspective of our theory in the next chapter.

To put our critical note more cautiously, it has remained an open problem to translate quantitative, deterministic theories in science into the framework of the INUS account. This account has certainly distinctive merits when we study qualitative causal relations in scientific and everyday contexts. And we should emphasize once more the commonalities between Baumgartner's refinement of the INUS account and our analysis. Both employ criteria of minimality and non-redundancy in order to exclude spurious causes. Both succeed for a wide range of causal scenarios.

Finally, a note on the ranking-theoretic analysis by Spohn (2006) is in order. This analysis aims to discriminate between spurious and genuine causes without using criteria of minimality and simplicity. Each possible world has a certain rank which is an inverse measure of its plausibility. The lower the rank, the more plausible the world is. Using ranks of possible worlds, Spohn defines what it is for a proposition to be reason for another proposition. Then the Humean convention comes into play to make the transition from reasons to causes. Our epochetic theory uses this transition as a template. It is indebted to Spohn's ranking-theoretic analysis for this reason.

However, the determination of the ranks of possible worlds relies on infor-

[11] See Bell (2023) for an overview of infinitary logics.

mation about the causal graph of the respective causal scenario. For example, Spohn (2006, p. 106) speaks of a 'conjunctive fork in which the ranks again just count the violations of the causal relations'. Reference to 'causal relations' must be understood here in the sense of genuine causal relations. If, by contrast, violations of spurious causal relations had the same impact on the ranks of possible worlds, Spohn's analysis would fail to work.

Some information about the genuine causal relations is therefore taken as primitive for the delineation between spurious and genuine causes. Since a ranking function plays the role of an epistemic state, it is assumed that the respective agent has information as to which possible worlds score well in conforming to the genuine causal relations. To give an abstract example: if *C* is a genuine cause of *E*, then a possible world in which *C* occurs without *E* violates a genuine causal relation. But not so if *C* is a mere spurious cause of *E*. In essence, the rank of a possible world serves as a measure of the violations of the genuinely causal laws. The ranking-theoretic analysis thus departs—perhaps unintentionally—from the Humean project of analysing causation without taking any causal notions as primitively given.

Our epochetic theory uses an ordering of epistemic priority too. The ordering is simple and characterized without causal notions. The first level contains laws, where the notion of law is understood in a minimalist and syntactic sense. The second level contains beliefs about presumed facts. Moreover, we assume that the beliefs of an epistemic state contain information about the temporal relations among events. Nothing less, nothing more.

Chapter 9

Simultaneous Causation

If the cause is simultaneous with its effect, how can we account for the direction of causation? In such cases, we may not find any asymmetry merely by looking at the given pair of cause and effect. However, we can recognize an asymmetry when we look for further explanations why cause and effect occurred, respectively. In essence, we are able to causally explain the occurrence of the cause in a manner which is independent of the simultaneous effect and in accordance with the Humean convention. But not the other way around. Simultaneous causation will thus be analysed in terms of Humean causal relations, in which the cause precedes its effect. This analysis draws on ideas about simultaneous causation in Dummett (1954).

1 Some Examples

Let's first look at some scenarios which have been adduced to show that there is simultaneous causation:

(1) A locomotive is pulling a coach.

(2) The rotation of a gear ring causes another gear ring to rotate.

(3) Lowering one end of a sea-saw causes the other end to go up.

(4) A lead ball is resting on a cushion, and so causes a deformation of the cushion.

(5) A force acting upon a body causes this body to accelerate.[1]

In some of these scenarios, we may be able to recognize a very small temporal delay between cause and effect. Since no material is absolutely rigid, the other end of the sea-saw does not immediately move up when we force one end to move down. Since gear rings made of steel are not absolutely rigid, rotation and momentum are not immediately transmitted among gears. We will not pursue this strategy of eliminating presumed instances of simultaneous causation for the following reasons.

Our theory of causation is aimed at capturing causal judgements. Such judgements are relative to an epistemic state and a model of the respective causal scenario. If we took into account that even sea-saws and gear rings made of steel are elastic and deformable, the resulting models would become very complicated. Another problem is that—in the above causal scenarios—we seem to be able to differentiate between cause and effect without recognizing small temporal delays between the two. Finally, we cannot recognize any temporal delay between total forces and corresponding accelerations in classical mechanics.

Despite our interest in simultaneous causation, it is worth noting that modern physics has eliminated simultaneous causal relations to a large extent. Relativistic theories in physics tell us that all fundamental interactions in physics are constrained by the speed of light. Roughly, an object exerts a force on another object in such a manner that the effect occurs after the time interval which light takes to travel from one object to the other. The details are complex, which is why we do not give an overview of relativistic physics here.

Kutach (2013) has delivered an account of causation in fundamental physics in which the asymmetry of causation is strictly aligned with the arrow of time. A cause always precedes its effect on his account. Backward and simultaneous causal relations are excluded. The argument against backward causation makes use of interventionist considerations, and seems to be a variant of the bilking argument to be discussed in the next chapter. Kutach shows his theory of causation to apply to classical gravitation, relativistic electromagnetism, general relativity, and quantum

[1] These examples are adopted from Huemer and Kovitz (2003), who in turn draw on Taylor (1966, p. 35), Brand (1980, p. 138), and Kant (1781/1998, A 203).

mechanics. Classical gravitational interactions are interpreted as arbitrarily fast instead of instantaneous.

Our envisioned theory of causation is, of course, in principle consistent with an account of causation in fundamental physics which avoids simultaneous and backward causation altogether. If there was only forward causation in fundamental physics, this would be great news for any reconsideration of the Humean convention, such as ours. We include simultaneous causation to make our theory more comprehensive and more easily applicable to causal scenarios where a temporal delay between cause and effect is difficult to recognize.

2 Causal Explanatory Asymmetries

In cases of simultaneous causation, we may not find any asymmetry merely by looking at the given pair of cause and effect. However, we seem to find an asymmetry when we look for further explanations why cause and effect occurred, respectively. In essence, while the occurrence of the cause can be causally explained independently of the effect, no such independent causal explanation is available for the occurrence of the effect. This account of simultaneous causation goes back to Dummett:

> If, then, the immediate cause is always simultaneous with its effect, how do we decide which of two events is the cause and which the effect? [...] *We determine which one is the cause by deciding which one can be already causally accounted for without reference to the other.* This statement is not viciously circular; our system of causal explanations is constructed piecemeal, and it is only when we already have a causal explanation of the occurrence of one of two events, each the sufficient and necessary condition of the other, that we can decide which of the two we are going to regard as the cause of the other. (Dummett 1954, p. 30n, emphasis added)

Let us now further elaborate this approach to simultaneous causation within our inferential framework. Deviating from Dummett's account, we will not refer to necessary and sufficient conditions in order to analyse simultaneous causal relations.

Recall that we identify causal relations in terms of inferential pathways between cause and effect. For A to be a cause of E, it is necessary that there is an inferential path from A to E which is forward-directed and which avoids redundant laws. In other words, $A \gg_{FN} E$ is a necessary condition for A to be a cause of E. Let's say that A explains E, possibly causally, iff $A \gg_{FN} E$, and both A and E are actual events. Using this simple and preliminary understanding of an inferential explanation, we can be more precise about explanatory asymmetries between two events. First, we define the notion of an explanatory inferential path from an event A to another event E such that this path does not go via a third event B.

Definition 27. $A \backslash B \gg_{FN} E$
Suppose S is an epistemic state such that A, E, and $A \gg_{FN} E$ are believed. We say there is an explanatory inferential path from A to E which does not go via B—and write $A \backslash B \gg_{FN} E \in K_{>}(S)$—iff there is a deduction of E from A which meets all conditions of Definition SRT_{FN} and the additional condition that B does not occur as an intermediate conclusion or premise. Nor does any sentence synonymous to B occur in this deduction.

In other words, $A \backslash B \gg_{FN} E$ means that there is an epistemic state S' such that (i) S' is uninformative on A and E, (ii) S' results from a contraction of S by $A \vee E$, and (iii) there is a deduction of E from $\bigcup S' \cup \{A\}$ such that this deduction is forward-directed, no law is redundant in $\bigcup S'$, and B does not occur as an intermediate conclusion or premise. Note that $A \backslash B \gg_{FN} E$ may be understood as a ternary conditional.

Why do we require that the inferential path from A to E must not use a sentence which is synonymous to B? Take the cushion which is deformed by a lead ball. We can also describe the event of placing a lead ball on the cushion by saying that a lead spherical object has been placed on the cushion. But we shouldn't say that there is an explanatory inferential path from placing a spherical object on the cushion to the deformation of the cushion which is independent of the event that a lead ball has been placed on the cushion. Admittedly, some vagueness is involved in the notion of synonymy. We can delimit, if not avoid, problems of synonymous descriptions by restricting the vocabulary in which the beliefs of the respective epistemic state may be expressed

Using the conditional $A \backslash B \gg_{FN} E$, we can say what it means that one event has explanatory priority over another.

Definition 28. Explanatory priority
Let S be an epistemic state. A and B are two events which are simultaneous, and believed to occur in the epistemic state S. All causal relations are understood in the sense of Definition 26. We say A is explanatorily prior to B iff the following two conditions are satisfied.

(1) There is C such that C causes A, and $C \backslash B \gg_{FN} A \in K_{>}(S)$.

(2) There is no C' such that C' causes B, and $C' \backslash A \gg_{FN} B \in K_{>}(S)$.

In brief, A is explanatorily prior to B iff it has a causal explanation independent of B, while B has no causal explanation independent of A. This relation of explanatory priority does not rest on any temporal relation between the events A and B. It rather concerns the causal explanations available for A and B, where causation is understood with the Humean convention in place and according to our analysis in the previous chapter. We are now in a position to reformulate Dummett's account of simultaneous causation in our framework.

Definition 29. Simultaneous cause
C is a simultaneous cause of E—relative to an epistemic state S—iff all of the following conditions hold:

(1) $C, E \in K(S)$

(2) $C \gg_{FN} E \in K_{>}(S)$

(3) C does not precede E, nor does E precede C

(4) C is explanatorily prior to E.

Recall that we have defined the notion of forward-directed deduction—which underlies the conditional $\gg_{FN}$—in the weak sense that no inferential step is backward-directed in time. For this reason, there are forward-directed deductions of an effect from a simultaneous cause. Understanding simultaneous causal relations in this way allows us to generalize our analysis of causation set forth in the previous chapter:

Definition 30. Cause
Let C and E be events. C causes E—relative to an epistemic state S—iff

(1) $C, E \in K(S)$,

(2) $C \gg_{FN} E \in K(S)$, and

(3) C precedes or is explanatorily prior to E.

This is our final reductive analysis of causation in Part II. It is reductive for the following reasons. Most notably, the analysis does not rely on causal models with structural equations. No causal relations are taken as antecedently understood or primitively given. While the relation of explanatory priority is defined in terms of causal relations, such relations are explained by our Humean analysis from the previous chapter. Notice, furthermore, that the analysis has the logical form of an explicit definition. Specifically, we can replace the relation of explanatory priority by its definiens. Likewise, we can replace reference to causal relations in the definition of explanatory priority by the analysans of our Humean analysis from the previous chapter. The latter analysis is reductive.

In the next chapter, we move on to backward causation, which may be seen as one of the three major challenges for a broadly Humean approach to the direction of causation. As previously indicated, the notion of backward causation has remained controversial in the literature. It is, in particular, controversial how the direction of causation may be understood in instances of backward causation. We will consider one specific account, and outline how backward causation may be integrated into our broadly Humean analysis. However, we will not endorse this integration for the time being. We are hesitant for mainly two reasons. First, there remain open problems for the specific account of backward causation considered. Second, the account implies that, so far, we lack empirical evidence for the claim that our world exhibits instances of backward causation. The above analysis is therefore our final reductive analysis of causation in Part II.

3 Causal Scenarios

Our analysis is now able to capture all scenarios of simultaneous causation looked at in the first section. Ad (1): the locomotive is pulling a coach. We read this statement in a causal manner. The motion of the locomotive causes the coach to move, but not vice versa. To apply our analysis, we need to show, first, that we can causally explain the motion of the locomotive independently of the motion of the coach. Second, we need to show that any causal explanation of the motion of the coach goes via the motion of the locomotive.

Why is the locomotive moving? It may be driven by a combustion or an electric engine. Suppose we have an old-fashioned steam engine, which is of the combustion type. The application of our analysis to other combustion engines and an electric engine is largely analogous to what follows. A steam engine is roughly based on the following chain of processes:

(1) Coal is burnt so that the temperature of a water reservoir reaches and stays at boiling point.

(2) When reaching boiling point, water evaporates.

(3) The evaporated water increases the pressure on a piston.

(4) The increased pressure on the piston makes the piston move.

(5) The motion of the piston is transmitted to cranks.

(6) The cranks drive the wheels of the locomotive.

(7) The locomotive moves because of the motion of the wheels.

(8) The coach is in motion because of the motion of the locomotive and the coach being attached to the locomotive.

Each of these processes is in itself causal. To be more precise, each process consists of two subprocesses, or events, such that one causes the other. Obviously, we have used broadly causal language to describe the processes. Speaking of transmission and saying that some motion is driving another motion indicate a causal direction. Also, the conjunctions *because* and *so that* are used with a causal meaning. Note that the processes form a sequence such that each element is a cause of its successor. Thus, we have a sequence of causal processes. Note that each process qualifies as an event which is temporally extended.

How can we justify the presumed causal directions in our description? Some of the causal connections seem to be of the simultaneous type. Assuming an absolutely rigid connection between the piston and the cranks, there is no temporal delay between the motion of the piston and the corresponding motion of the cranks. Likewise for the motion of the cranks and the motion of the wheels. However, at least for the first two elements in our sequence, we can justify the presumed causal direction by the Humean convention. The temporally extended event of burning coal begins at a time before water reaches boiling point. Coal needs to be burnt first before water evaporates.

To see the causal order between the first elements of the sequence more clearly, let's assume for contradiction the causal hypothesis that water reaching and staying at boiling point causes coal to burn in a steam engine. If we make this assumption, we need to look for a causal explanation of the water reaching and staying at boiling point. Such an explanation is needed since the event of water reaching and staying at boiling point—in an environment which is colder than the boiling point of water—is not an event sui generis. It is not caused by itself, and therefore requires a causal explanation. No such explanation can be found. Specifically, the motion of the piston cannot be used in a causal explanation of the temperature of water. For the piston does not move before the water reaches boiling point and evaporates. Again, the Humean convention proves surprisingly powerful: a causal explanation of the temperature of water by the motion of the piston amounts to a violation of this convention. And while it is true that the piston may be used to increase the temperature and pressure of air, and thereby increase the temperature of water, this is not what's happening in a steam engine. For evaporated air is constantly released by the steam engine. This is why we are seeing white smoke coming out of a steam-driven locomotive.

It may also be worth noting that burning coal is an irreversible process since it goes along with an increase of entropy. The second law of thermodynamics tells us that the entropy of a closed system never decreases, at least not at the macroscopic level. A locomotive driven by a steam engine, taken together with the surrounding air into which the steam is released, is considered a thermodynamically closed system. While it has become technologically feasible to reduce carbon dioxide to carbon, such processes are not simple reversals of combustion. Specific catalysts and conditions are needed to obtain carbon from carbon dioxide. By contrast, the rotation of cranks and wheels are reversible processes. Wheels and cranks can rotate both forward and backward without any violation of the second law of thermodynamics. Sometimes a locomotive goes backward.

The crucial point is that we can use our knowledge about the causal order of the initial elements in the sequence of processes in order to assign a causal order to the others. For we have a Humean causal explanation of the rotation of the cranks which begins with the process of burning coal and goes via the increase of water temperature, the evaporation of water, the increase of pressure on the piston, and the motion of the piston. This explanation can be extended to a Humean causal explanation of the motion of the coach by adding three more elements: the rotation of the wheels of the locomotive, the motion of the locomotive, and the motion of the coach. By a Humean causal explanation, we mean one where the cause precedes the effect—at least in the sense that the temporally extended cause begins earlier than the temporally extended effect. Clearly, while in a running steam engine coal is being burnt at the same time at which the cranks are rotating, the process of burning coal started earlier than the rotation of the cranks.

By contrast, we do not have a Humean causal explanation of the rotation of the cranks which avoids reference to the motion of the piston, increase of pressure on the piston, evaporation of water, and heating of water by burning coal—provided the locomotive is not going downward, is not pushed by another locomotive, etc. Put more carefully, looking at standard accounts of the key processes in a steam engine, we cannot identify an event C' which satisfies the following two conditions: first, C' precedes the rotation of the cranks and the motion of the coach, but is different from burning coal, water evaporating, and the increase of pressure on the piston. Second, we can infer the rotation of the cranks and the motion of the coach from C' in a forward-directed manner such that no law is redundant.

We have thus shown that (i) the process of burning coal stands in the relation of explanatory priority to the rotation of the cranks, the motion of the locomotive, and the motion of the coach. Moreover, (ii) there is a forward-directed inferential path from burning coal to the rotation of the cranks, the motion of the locomotive, and the motion of the coach. (iii) The laws of this path are non-redundant, as will be shown shortly. (iv) The rotation of the cranks and wheels of the locomotive as well as the motion of the locomotive itself are simultaneous to the motion of the coach, given these parts are connected absolutely rigidly. If the connections are not fully rigid, a Humean analysis of the causal relations in question can be even more easily justified, as we have seen in Section 1. (i) to (iv) imply that the motion of the locomotive is a simultaneous cause of the motion of the coach on our analysis of simultaneous causation. Likewise, the rotation of the cranks is a simultaneous cause of the rotation of the wheels, and the latter is a simultaneous cause of the motion of the locomotive.

The sceptical reader may want to see a justification for claim (iii). The inferential path in question may be established by macroscopic laws from engineering and phenomenological thermodynamics. These laws are redundant iff they can be derived from more fundamental laws of physics given by the background theories in the respective epistemic state. Now, we have to distinguish two cases. First, the macroscopic laws in question are non-redundant relative to the epistemic state considered. Then the inferential path with the macroscopic laws is free of redundant laws. Second, the macroscopic laws are redundant since they can be derived from more fundamental laws. Then, however, we can transform the inferential path with the macroscopic laws into a path with the more fundamental laws without changing the temporal relations among events. The inferential path thus obtained is free of redundant laws. Either way, claim (iii) holds true.

We have also shown that the motion of the coach does not stand in the relation of explanatory priority to the motion of the locomotive. There is no cause of the motion of the coach whose explanatory, inferential pathway does not go through the motion of the locomotive. The motion of the coach is therefore not a simultaneous cause of the motion of the locomotive. Likewise, the rotation of the wheels is not explanatorily prior to the rotation of the cranks. Hence, the former is not a simultaneous cause of the latter.

Ad (2) in Section 1: the rotation of a gear ring causes another gear ring to rotate. This causal scenario can be dealt with in a manner which is analogous to the preceding example of a locomotive. Without loss of generality, think of a bicycle where two sets of gear rings, also called *sprockets*, are connected by a chain. To make our problem harder, let us assume the chain is absolutely rigid. Suppose you are riding your non-electric bicycle, and the road goes upward. There is no strong tailwind. Nobody is pushing your bicycle upward. Then we can roughly describe the dynamics of the bicycle by the following chain of events:

(1) Your legs exert a force on the pedals.

(2) The two cranks rotate because of this force on the pedals.

(3) The rotation of the cranks is transmitted to the front gear ring.

(4) The rotation of the front gear ring is transmitted to the rear gear ring by the chain.

(5) The rear wheel rotates because of the rotation of the rear gear.

(6) The whole bicycle is set in motion by the rotation of the rear wheel.

As in the previous scenario, this description assumes a certain causal order such that each element is a cause of its direct successor. Speaking of *exertion of forces* and *transmission of a rotation* clearly indicates a causal direction. The use of *because* in (2) and the conjunction *by* in (6) have a causal meaning. How can we justify the presumed causal order? Why do we say that our legs exert a force on the pedals rather than the other way around?

Again, there is a combustion engine in the background. Our legs move and exert a force on the pedals because of contraction of certain muscles. This contraction, in turn, is driven by exothermic biochemical reactions which go along with an increase of entropy. The ultimate source of energy in human cells is the oxidation of carbohydrates, fats, and amino acids. Contraction of muscles is more directly driven by hydrolysis of ATP (adenosine triphosphate), where ATP is generated by oxidation of carbohydrates. The chemical pathways are complex and sophisticated, much more sophisticated than our latest technology used in combustion engines.[2]

Most notably, hydrolysis of ATP precedes the contraction of muscles, just as some fuel needs to be burnt first before a man-made combustion engine is driving any wheel or crank. In other words, contraction of muscles doesn't happen simultaneously to the underlying exothermic biochemical reactions. Hence, we have a causal Humean explanation of contraction of muscles and rotation of the cranks where the cause precedes the effect. This causal explanation can be extended to a causal explanation of the rotation of the front gear ring since there is a rigid connection between cranks and the front gear ring. Rotation of cranks and front gear ring is simultaneous, given the connection between the two is absolutely rigid. By the same pattern, the causal explanation of the rotation of the front gear ring can be

[2] Any comprehensive undergraduate textbook of biology may be used for an overview of the biochemistry underlying contraction of muscles.

extended to a causal explanation of the rotation of the rear gear ring. The rotation of the two gear rings happens simultaneously.

By contrast, standard descriptions of how a bicycle works contain no process which could serve as a causal Humean explanation of the rotation of the rear gear ring other than the one which begins with contraction of muscles and hydrolysis of ATP. To be more precise, in the absence of en electric engine, strong tailwind, a downward gradient, etc., there is simply no causal explanation of the rotation of rear gear ring which avoids reference to some human legs pedalling the bicycle. To give a simple illustration, suppose you are cycling upward a steep pass in the mountains. No electric engine or fellow human is helping you. Then it's only your muscles which drive your bicycle with all its gear rings. Hence, given the specific assumptions of the causal scenario, the rotation of the front gear ring stands in relation of explanatory priority to the rotation of the rear gear ring. Hence, the former is a simultaneous cause of the latter.

Of course, the transmission of motion can also go from the rear to the front gear. Suppose your bicycle doesn't have a freewheel. Colloquially, such a fixed-wheel bike is called a *fixie*. Now suppose you walk your fixie on the pedestrian way. Then rotation of the rear wheel forces the rear gear to rotate. The latter rotation, in turn, is transmitted to the front gear and its crank via the chain. In this case, we have a causal Humean explanation of the rotation of the rear gear which is independent of the rotation of the front gear. Unlike when you pedal your bicycle, the rear gear causes the front gear to rotate. The Humean cause of the rotation of the rear wheel with its gears is given by the contraction of muscles which exert a force on the bicycle as a whole since we have now assumed that you are walking your bicycle. On this assumption, there is no causal Humean explanation of the rotation of the front gear independent of the rotation of the rear gear. Hence, this time, rotation of the rear gear stands in the relation of explanatory priority to the rotation of the front gear. The former is therefore a cause of the latter.

Ad (3) in Section 1: lowering one end of a sea-saw causes the other end to go up. Using our expertise about biochemical combustion engines and bicycles, we can be briefer about the sea-saw. Suppose some human pushes one end of the sea-saw down so that the other end moves up. Obviously, pushing one end of a sea-saw down requires contraction of muscles. This contraction, in turn, is driven by certain biochemical exothermic reactions which set on very slightly before any contraction happens. Hence, we have a Humean cause of the downward movement of one end of the sea-saw which is independent of the upward movement of the other end. The reverse, however, is not true: we don't have a Humean cause of why one end of the sea-saw moves upward which is independent of the downward movement of the other end. For this reason, the former isn't on a par with the latter. Pushing one end of the sea-saw down is thus shown to be cause of the upward movement of the other end on our analysis.

Ad (4): a cushion is deformed by a lead ball resting on it. The wording of this sentence seems to imply that the lead ball causes the deformation. However, we could also argue that the position of the ball is caused by the cushion. There is a downward, gravitational force of the lead ball which is countered by the force of the resistance of the cushion. The deformation of the cushion is inelastic. Again, we have to look for further causal explanations. Notably, a lead ball doesn't come out of the blue. Given presumed laws of nature, there must have been some force which placed the ball on the cushion. For example, someone took the ball in her hand and put it there. The force exerted by the person's hand and fingers explains why the lead ball ended up on the cushion. It is caused by contractions of muscles, which are driven by exothermic biochemical reactions. Similar considerations apply to causal scenarios where a lead ball has been placed on the cushion by a robot.

Ad (5): forces cause a body to accelerate. In Aristotelian physics, any motion of a body requires a force. In Newtonian physics, it's only acceleration of a body which does. The link between force and acceleration is commonly understood as causal in nature, even though there is no temporal delay between the two.[3] For clarification, we need to distinguish between total forces and special forces, such as gravitational and electromagnetic forces. The total force acting upon a body is the vector sum of the component forces acting upon it. Newton's second equation tells us that the vector of the total force equals the product of mass and acceleration:

$$\mathbf{F} = m \cdot \mathbf{a}.$$

F stands for the force vector, *m* for the mass of the accelerated object, and **a** for the acceleration. The acceleration happens simultaneously to the acting of the total force. Newton's second equation is nonetheless standardly

[3] See, e.g., Huemer and Kovitz (2003).

interpreted in a causal manner. This interpretation is suggested by the connotations of the Latin word *vis*, the English word *force*, the French word *force*, etc. The causal interpretation is also driven by our desire to establish causal explanations of spatiotemporal phenomena. It's less obvious whether we should also understand the determination of total forces by component forces as causal, though.

As with previous examples, we need to look for causal explanations of total forces and accelerations, respectively, in order to recognize some explanatory asymmetry between the presumed cause and its effect. Total forces are determined by special forces, which in turn are determined by specific circumstances. The gravitational force between two bodies, for example, is determined by their masses and the distance between the two according to the following equation:

$$\mathbf{F}_{1,2} = -\gamma \cdot \frac{m_1 \cdot m_2}{|\mathbf{r}_1 - \mathbf{r}_2|^3} \cdot (\mathbf{r}_1 - \mathbf{r}_2).$$

$\mathbf{F}_{1,2}$ stands for the gravitational force which is exerted on the body with mass m_1 by the body with mass m_2. $\mathbf{r}_1$ and $\mathbf{r}_2$ stand for the positions of the two bodies in space, represented by vectors. γ is the gravitational constant. When merely doing calculations, we can use this equation to determine various quantities: masses, forces, and positions of objects. At the same time, there is some understanding that gravitational forces are actually determined by the masses of and the distances between bodies, but not vice versa. It seems wrong to think that a gravitational force causes the mass of an object. Likewise, we don't think that spatial positions of bodies are directly caused by simultaneous forces, even though forces cause accelerations, which in turn affect the spatial positions of bodies.

Again, the Humean convention—broadly construed so as to apply to temporally extended intervals—helps distinguish between causes and effects with respect to Newton's law of gravitation. In classical mechanics, the mass of a body is constant, given the body remains intact. The temporally extended event of a body having a certain mass therefore precedes the event of this body exerting and receiving a certain gravitational force at a certain time—in the sense that the starting point of the former event precedes that of the latter. Recall that we have refined the Humean convention for temporally extended events in Section 5.

Things are less straightforward for the relation between spatial position and gravitational force since spatial position is not constant in classical mechanics. However, the spatial position s of an object at time t is approximately determined by its spatial position s' at a previous point t' in time and its velocity at t', given the temporal distance between t and t' is 'small'. The margin of error of this determination can be rendered infinitesimally small by assuming the distance between t and t' to be arbitrarily small. Hence, at least at the level of infinitesimals, we have a causal Humean explanation of the position of a body in classical mechanics. Position and velocity cause the position of an object at an infinitesimally later time point in the sense of our Humean approach to causation in Chapter 7: for C to be a cause of E, C must precede E, and there must be a forward-directed inferential path from C to E such that any law of this path is non-redundant. The latter deduction uses beliefs which we can retain when suspending judgement on C and E.

These considerations show that there is a forward-directed inferential path from the mass of a body and its spatial position at time t' to the gravitational force at time t such that two conditions are satisfied. First, t is infinitesimally later than t'. Second, the temporally extended event of having a certain mass begins at a time prior to t. It's easy to show that all laws of the inferential path in question are non-redundant. Provided we know 'enough' about other forces acting upon the two bodies of our system, we can extend the inferential path to one which goes to the total force acting upon these bodies. Hence, we have a Humean causal explanation of the total forces acting in our system. By contrast, there is no forward-directed inferential path that determines the accelerations of our bodies at time t such that this path does not go via the total forces at time t. Hence, the event of there being a certain total force stands in relation of explanatory priority to the event of a body being accelerated according to classical mechanics. Total forces therefore cause accelerations, but not vice versa, on our analysis of simultaneous causation.

What about forces other than those of classical gravitational interactions? As mathematicians would put it, we claim without proof that a similar demonstration can be given for other types of forces, such as electromagnetic and contact forces. That is, we can give causal Humean explanations of such forces in terms of specific circumstances, such as the charges and positions of objects. Needless to say, the inferential determination of total forces by component forces is always relative to a specific model of the physical system considered. There is no way to make the qualification that we need to know 'enough' about all the component forces acting on an object—in order to determine the total force—entirely precise. For real-world systems, it's almost always impossible to consider all component forces. Arguably, this problem disappears only from the hypothetical viewpoint of omniscience.

Do the specific component forces acting on a body at a specific time cause the total force acting on this body at that time? We find it difficult to form an intuition here. The vector sum of the component forces acting on an object seems a mere mathematical auxiliary, which makes the notation of the formalism easier. However, on our analysis of simultaneous causation specific component forces qualify as causes of total forces. In the absence of firm intuitions, this result seems at least not counterintuitive. Experts on grounding might prefer to view the relation in question as one of grounding since a component force is part of a total force. However, our intuitions concerning the distinction between grounding and simultaneous causation are not always clear-cut. Fine (2012, p. 12), for example, thinks that total forces ground corresponding accelerations. This judgement deviates from the literature on simultaneous causation and the everyday meaning of the notion of force.

Suppose we want to maintain a strict distinction between simultaneous causation and grounding. This may be achieved by a relatively simple constraint on causes and their effects in the above account of simultaneous causation. We simply require that the event of the cause is distinct from the event of the effect. To make this notion precise, let us exploit the observation that all events concern a spatiotemporal region. No event seems to be outside of space and time. So we suggest that two events are distinct iff the two spatiotemporal regions—corresponding to these two events—do not stand in a relation of identity or containment to one another. Note that the scenarios (1) to (4) of simultaneous causation all satisfy this constraint. The presumed causes are distinct from their effects in the sense just explained. Things are less clear-cut for scenario (5).[4]

There remains to discuss briefly a causal scenario considered by Hausman (1998, 44n). Suppose, you take two boards and position them in such a

[4] The suggested explanation of when two events are distinct is akin to the one by Lewis (1986a).

manner that they lean against one another. Each board would fall if it was not leaning against the other. Hausman argues that this is a scenario of simultaneous causation. The position of each board is caused by the other. But we can also view this causal scenario as a case of prevention: the position of each board prevents the other from falling to the ground. The event of falling down, which is prevented, is in the future of the event of resisting the other board. Hence, we have at least an interpretation of the scenario which conforms to the Humean convention. In addition, we have a Humean causal explanation why the boards are where they are. I have put them at their current position with my hands, and this action precedes the extended event of the boards occupying a certain position in space.

4 The Length of the Pendulum

Our account of simultaneous causation may be used to shed some light on a few more causal scenarios which received a great deal of attention in the literature on causation and explanation. In this section, we look at the tower-shadow asymmetry, the length of the pendulum, and the ideal gas law. Both the tower-shadow asymmetry and the pendulum have been adduced to show limitations of the old DN-account of explanation (see, e.g, Bromberger (1966) and Woodward (2003, Ch. 4)). We have a deductive nomological inference from the length of the shadow to the height of the tower, but intuitively we are hesitant to view this inference as explanatory. Likewise, we can approximately derive the length of the pendulum from its period by the following equation:

$$T = 2\pi\sqrt{\frac{l}{g}}.$$

T stands for the period, l for the length, and g for the acceleration of freely falling bodies at the surface of the Earth. But we are hesitant to say that the length of the pendulum can be explained by its period. Woodward (2003), consequently, uses these and other counterexamples to the DN model of explanation to argue for a broadly causal account of explanation. Hempel (1965, p. 352), by contrast, argued that an inferential explanation of the period of a pendulum in terms of its length is not of the causal type because no law of succession is involved.

Let's begin with the tower-shadow asymmetry. Obviously, if we take into account that the speed of light is finite, the Humean convention suffices to recognize the tower as cause of its shadow. But even in the absence of this information, we are now in a position to recognize some explanatory priority between the tower and its shadow. For we have a causal Humean explanation of the height which is independent of the shadow. For example, the height of the tower may be explained by the fact that such and such a number of layers of bricks have been built upon one another at a time prior to the casting of the shadow.

One may object to the presumed causal nature of this explanation that the event of bricks forming a number of layers doesn't precede the event of the tower having a certain height. Once the last layer of bricks had been laid and fixed by cement, the tower began to have a certain height. However, there is a causal Humean explanation of how the bricks came to form layers in the first place. Someone simply used his hands and tools to build the structure of bricks. Combustion engines, broadly conceived as in the previous sections, did some work to build the structure. And these engines started their work before the tower with its height came into being.[5] By contrast, there is no causal Humean explanation of the length of the shadow independent of the height of the tower. Hence, the height of the tower stands in the relation of explanatory priority to the length of the shadow.

Similar considerations apply to the length of the pendulum in relation to its period. Think of a metronome used for the training of musicians to mark musical tempo. At least a few decades ago, a metronome consisted of a pendulum whose length could be adjusted manually. This adjustment satisfies the Humean convention: someone used her hands to fasten a movable weight so that the pendulum is set to a certain length. And these actions started at a time which precedes the event of the metronome being set to a certain length.[6] If the pendulum is rigid such that the length cannot be adjusted, then it has some causal history of how it was manufactured to have such a length. A pendulum needs to be manufactured before it can be made to swing back and forth, and so the Humean convention is

[5] Since the layers of bricks are parts of the tower, they may also be said to ground the tower which has a certain length.

[6] Without loss of generality, our account of a metronome ignores the mass of those parts of the physical pendulum which are different from the movable weight. Ignoring the mass of those parts may also be justified as an idealization because the mass of the movable weight is much greater than that of the other parts of the pendulum.

satisfied. Hence, we have a causal Humean explanation of the length of the pendulum which is independent of its period. By contrast, there is no such Humean causal explanation of the period of the pendulum which is independent of its length. Hence, the length of the pendulum stands in the relation of explanatory priority to its period.

One may object to this order of explanation by arguing that a metronome gives us a nice example where it makes perfect sense to say that the period of the pendulum explains its length. The musician adjusts the length of the metronome's pendulum in such a manner that it marks a certain musical tempo.[7] So, it's the tempo marked by the period of the pendulum which causes it to have a certain length, but not vice versa. This objection is based on a teleological explanation. Our reply to the objection is twofold. First, in line with wide-ranging consensus in philosophy and science, we think it desirable to avoid teleological explanations and causes as much as possible. Second, we prefer to say that it's the musician's intention to let the metronome swing with a certain period which causes a certain setting of the length. The temporally extended interval of having this intention precedes the manual adjustment of the length of the pendulum.

Let's move on to the ideal gas law. This law has two standard formulations:

$$\frac{p \cdot V}{T} = const$$
$$p \cdot V = R \cdot T \cdot n.$$

p stands for pressure of the gas, V for its volume, and T for its temperature. R is the gas constant and n a quantity called *amount of substance* which is a measure of the number of molecules. Hempel (1962, p. 12) thinks that the ideal gas law may be used to explain the value of some of its quantities in the static case where pressure, volume, temperature, and the number of molecules are constant. Woodward (2003, p. 234), by contrast, holds that the law helps us explain phenomena only in dynamic scenarios where some of its quantities are in flux. In line with Woodward, we find it difficult to see the ideal gas law doing interesting explanatory work in the static case. For we seem to be at a loss when asked to distinguish between explanans and explanandum. The problem is that each quantity can be inferentially determined using the set of the other quantities.

[7] See van Fraassen (1980, pp. 132–134) for a similar story concerning the height of a flagpole and its shadow.

Assuming our account of simultaneous causation, we show that the ideal gas law is rather used in causal explanations of *changes* of pressure, volume, and temperature, respectively. Taking this law at face value, we have to say that a change in one quantity goes along with immediate changes of some of the other quantities. It's difficult to recognize any temporal delay between, say, a change in temperature and the changes of pressure and volume. However, we can recognize external forces and events which cause directly a change of one quantity without directly causing also the changes of the other quantities. Think of a steam engine, discussed in the previous section. Water evaporates and so the number of molecules in the gaseous state increases by a large amount. Then we can use the ideal gas law to calculate the impact on the product of pressure and volume. Clearly, the increase in the number of gas molecules stands in the relation of explanatory priority to the changes of the other quantities. For there is a Humean causal explanation of this increase in terms of heating water and its subsequent evaporation. Water needs to be heated before it evaporates. It's not the increase in pressure in the cylinder which causes the water to evaporate.

It's easy to adduce further examples of causal explanation using the ideal gas law. Think of pumping air into the tyres of your bicycle. Thereby, the number of molecules in the tyre increases, which leads to an increase of pressure and volume. If we deflate a properly inflated tyre, the number of molecules in the tyre decreases, which causes pressure and volume to decrease. In either case, there is an external cause which changes the quantity n (amount of substance) and which precedes the event of the number of molecules being set to a certain value. At least the activation of our muscles which lets us inflate or deflate a tyre precedes any change in the number of molecules in the tyre.

Arguably, our examples of causal explanation using the ideal gas law are instances of simultaneous causation: even though the change of one quantity is simultaneous to that of another, we can identify direct causal explanations only for changes of a certain quantity, but not for all. Which change can be externally and thus causally explained—external to the gas itself—varies among causal scenarios of thermodynamics.

Chapter 10

Backward Causation

Causation seems to be an asymmetric relation: if C is a cause of E, then E is not a cause of C. Causal scenarios with cycles are rare to find and controversial. In any case, causation is directed in the sense that we think cause and effect have different roles. Intuitively, causes bring about their effects, but not vice versa. How can we distinguish causes from their effects?

The Humean convention gives a simple answer to this question: a cause always precedes its effect in time. This account works surprisingly well for a wide range of causal scenarios, particularly in everyday contexts. However, we must wonder if backward causation is at least a conceptual possibility. What could it mean that this type of causation is conceptually possible? Should we dispense with the Humean convention to make room for backward causation?

Abandoning the Humean convention completely seems unreasonable in view of rather severe problems with the notion of backward causation and alternative approaches to the direction of causation. In this chapter, we will review such problems. Thereby, we will establish further results in favour of a reconsideration of the Humean convention. First, Woodward's interventionist test for causation implicitly relies on the Humean convention. If we abandon this convention, the interventionist test becomes indeterminate as regards the direction of causation. Second, the Humean convention plays an important role in Reichenbach's account of the direction of time and causation.

Third, Lewis's counterfactual approach to the direction of causation runs into seemingly insurmountable problems. We will add one more severe problem to the list of problems which have already been observed in the literature. In brief, we show that Lewis's fork theory implies that there is unlimited growth of concurrent events. Fourth, Hausman's independence principle does not suffice to characterize the direction of causation in time-symmetric theories. This problem arises for the refinement of the INUS account by Baumgartner and Falk (2019) as well.

We will also discuss the *bilking argument* against backward causation. The discussion reveals that empirical and theoretical arguments in favour of backward causation, if available at all, are indirect and open to interpretation to a higher degree than our reasoning is for other theoretical hypotheses in science.

In the final section, we will nonetheless outline how backward causation may be understood within our inferential and broadly Humean approach to causation. The proposal amounts to a disjunctive approach to the direction of causation, which goes back to Dowe (1996). On this proposal, the Humean convention remains one of two means to distinguish between causes and effects.

1 From an Interventionist Perspective

The Humean approach to the direction of causation is simple and straightforward: a cause always precedes its effect in time. Obviously, this excludes the possibility of backward causation. Interventionist approaches to causation seem to be more liberal in this regard. Roughly, if there are interventions on variable X which change the value of variable Y, then X is a cause of Y. If, by contrast, there is no way to change Y by intervening on X, then X cannot be such a cause. So, the conceptual possibility of backward causation seems to open up once we adopt an interventionist account of causation (Gebharter et al. 2019).

Contrary to this line of reasoning, Reichenbach (1956) and Black (1956) argued that an interventionist account of causation rules out the conceptual possibility of backward causation. In essence, Reichenbach argues that the outcome of an intervention is only well defined if we accept the Humean convention about the direction of causation. Black (1956) tries to refute the hypothetical assumption of backward causation by drawing on an interventionist understanding of causation. This refutation is often referred to as the *bilking argument* against backward causation. The intended meaning seems to be as follows: if there was backward causation, we could bilk the effect out of its cause. In other words, we could take away the cause from the effect after the occurrence of the latter. The effect could happen without its cause, and so the effect would be uncaused. This seems contradictory in a deterministic setting. Using Woodward's interventionist account of causation, we will give a precise formulation of the bilking argument.

The arguments by Reichenbach (1956) and Black (1956) have received relatively little attention in the more contemporary literature on causation.[1] This lack of reception is surprising since the interventionist account has gained unprecedented popularity among philosophers and other researchers in the wake of the seminal works by Woodward and Pearl. Neither Woodward (2003) nor Pearl (2009) explicitly address the problem of backward causation. In what follows, we explain the key elements of Woodward's interventionist account of causation. We then review the arguments by Reichenbach (1956) and Black (1956) against the background of this account.

2 The Interventionist Account with Causal Graphs

The interventionist account of causation is centred on a simple idea: X causes Y iff there is an intervention on X which changes the value of Y. What does it mean to intervene on a variable X? What does it mean that an intervention on the value of a variable makes another variable to change its value? Woodward (2003) develops a whole theory to answer these questions. This theory makes essential use of the framework of causal models by Spirtes et al. (1993) and Pearl (2000). Since we have made ourselves familiar with the basic concepts of causal models, we can get to the core of Woodward's analysis without much further preparation.

The proposed semantic analysis of causal relations rests on the notion of an ideal intervention: X causes Y iff there is an intervention I on the putative

[1] Price (1996, Ch. 7) and Dowe (1996) are notable exceptions. The bilking argument is given careful consideration there. Friederich and Evans (2019), Gebharter et al. (2019), and Hausman (1998, Ch. 12) include a brief discussion of this argument.

cause X which changes the putative effect Y such that—relative to some variable set—I fulfils a number of conditions which ensure that any change in Y following I is to be ascribed to X, and X only. I is an ideal intervention which tests whether X causes Y iff the following conditions are met:

> (**IV**)
>
> I1. *I causes X.*
>
> I2. I acts as a switch for all other variables that *cause* X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
>
> I3. Any directed path from I to Y goes through X. That is, I does not directly *cause* Y and is not a *cause* of any *causes* of Y that are distinct from X except, of course, for those *causes* of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.
>
> I4. I is (statistically) independent of any variable Z that *causes* Y and that is on a directed path that does not go through X. (Woodward (2003, p. 98, emphasis added))

The notion of intervention thus is a ternary relation: I intervenes on X relative to Y. The interventionist test for causation goes as follows:

> (**M**) X is a (type-level) direct cause of Y with respect to a variable set **V** iff there is a possible intervention on X with respect to Y that will change Y or the probability distribution of Y when one holds fixed at some value all other variables $Z_i \in \mathbf{V}$. X is a (type-level) contributing cause of Y with respect to **V** iff (i) there is a directed path from X to Y such that each link in this path is a direct causal relationship and (ii) there is an intervention on X that will change Y or the probability distribution of Y when all other variables in **V** that are not on this path are fixed at some value. (cf. Woodward (2003, p. 59))

This interventionist analysis of causation is not reductive, as Woodward himself is happy to admit. A great deal of information about the causal graph of the respective causal scenario is needed in order to apply—practically or theoretically—the interventionist analysis in question to a concrete causal claim. No assumptions, however, are made about causal relations between X and Y in the analysis of causal relations between X and Y. This is why Woodward thinks that his analysis is not viciously circular.

Woodward uses both indicative and subjunctive conditionals to describe what happens under an intervention. The indicative formulation may be useful for actual experiments and actual interventions: if we change the value of this variable, we will observe such and such changes of certain other variables. The subjunctive formulation seems more appropriate for what Woodward calls *hypothetical interventions* and *hypothetical experiments* (Woodward 2003, p. 128n). Interventions of hypothetical experiments do not have to be practically feasible. Even nomological feasibility is not a requirement. The notions of hypothetical intervention and hypothetical experiment therefore introduce a counterfactual element into the interventionist analysis. For the discussion to follow, it is important to note that Woodward himself advertises his interventionist analysis as a counterfactual one:

> I argue below that to elucidate certain kinds of causal claims, including claims about direct causal relationships and singular causal claims, one must appeal to counterfactuals with complex antecedents—counterfactuals that describe what will happen under combinations of manipulations or interventions, rather than under single manipulations. (Woodward 2003, p. 21)

If Woodward had merely aimed at a theory of actual experiments, there might be a way to understand (**M**) without counterfactuals. Since, however, his analysis is intended to have greater scope, (**M**) must allow for a counterfactual reading. This reading is furthermore implied by the leitmotiv of Woodward's book: we need to ask *what-if-things-had-been-different* questions in order to understand causal explanation. This is not to say that (**M**) is confined to counterfactual interventions. Both actual and counterfactual interventions are important to the interventionist account. We call an intervention *actual* iff it is actually carried it. Otherwise, the intervention is called *counterfactual* or *hypothetical*.[2]

In light of its counterfactual elements, we can view Woodward's interventionist account of causation as an attempt to improve on Lewis' counterfactual account by exploiting the resources of causal models. Causal claims are analysed in terms of interventionist counterfactuals rather than variably strict conditionals as defined in Stalnaker (1968) and Lewis (1973b). With some qualifications, the conceptual order remains the same as in Lewis's account: counterfactuals provide the semantic foundation of causal claims.

3 The Direction of Interventions

When determining counterfactual consequences, we think about scenarios in which things are different from what they actually are. Some things change, others remain the same. Which elements of a counterfactual scenario stay the same? Which elements are going to change? For the semantics of variably strict conditionals, Lewis (1979) gave some heuristic principles to answer these questions. Likewise, Woodward (2003) imposes certain constraints on interventions, explicated by (**IV**), in order to explain the semantics of interventionist conditionals. The question thus arises if these constraints suffice to yield a semantics in which the relevant interventionist conditionals have determinate truth values.

In his seminal *The Direction of Time* (1956, Ch. 6), Reichenbach devises an argument which may be used to show that the semantics of interventionist conditionals remains indeterminate without the Humean convention. His argument is based on a causal scenario along the following lines. Suppose there are two billiard balls on a billiard table. Let us name them a and b. a moves straight and unaccelerated toward b, which is at rest. Then a collides with b so that a changes the direction of its motion. Also, it pushes b to move. b is a bit larger and heavier than a. At time t_1, a is at place $\vec{x}_1$, at time t_2 it collides with b at place $\vec{x}_2$. Finally, a ends up being at place $\vec{x}_3$ at time t_3. Here is a graphical illustration:

[2] For further statements that make the counterfactual content of (**M**) explicit, see p. 10, 11, 17, and 57 in Woodward (2003).

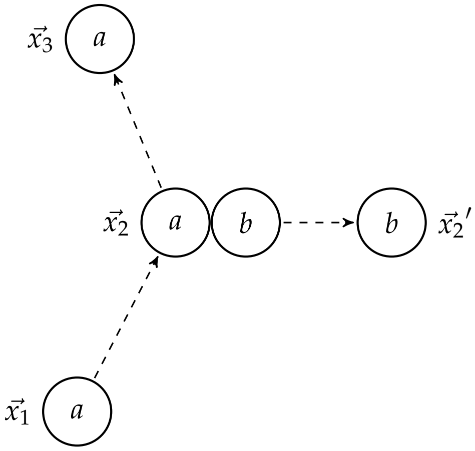


Figure 69: Collision between a and b at $\vec{x}_2$

We think of the collision as a cause of a being at place $\vec{x}_3$ at time t_3. This may be justified by a brief counterfactual consideration: had a not collided with b at t_2, it would not have been at position $\vec{x}_3$ at t_3. In this consideration, we keep the past of t_2 fixed, while the future of t_2 is open to changes. The collision may be represented by a binary variable C. We can intervene on C simply by taking out billiard ball b. Thereby, the value of C is set to false.

However, we must wonder why we do not hold the future of t_2 fixed. It seems as if the following counterfactual makes sense as well: had there been no collision between a and b, a would have come from another direction in the first place. So, the collision is a cause of a being at place $\vec{x}_1$ at time t_1. It is a backward cause of the latter event. Once we have liberated ourselves from the Humean convention, we can see a lot of backward causation in the world. The forward and the backward interpretation are, respectively, depicted by the following figures:

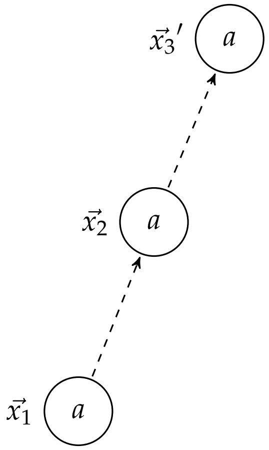


Figure 70: Counterfactual scenario 1: no collision at $\vec{x}_2$ and past is fixed

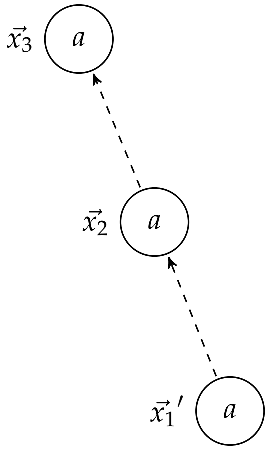


Figure 71: Counterfactual scenario 2: no collision at $\vec{x}_2$ and future is fixed

What's wrong with the backward interpretation? Which reasons do we have for holding the past of interventions fixed as opposed to the future? It seems as if an actual experiment could decide the matter. Let us take the billiard ball *b* from the billiard table. Thereby, we intervene on the trajectory of the moving billiard ball *a* so that no collision between the two balls occurs. Then we look at what happens when a billiard ball moves from $\vec{x}_1$ to $\vec{x}_2$ in a straight line, without acceleration, and with the same velocity as in the original setting. We observe that the ball does not end up at position $\vec{x}_3$ after the time interval $t_3 - t_1$. So, the collision is a cause of the ball *a* being at position $\vec{x}_3$ at time t_3.

However, the proponent of the backward interpretation will not accept the setup of the experiment. She looks at billiard balls moving straight from $\vec{x}_2$ to $\vec{x}_3$ since she thinks we should hold the future of the collision fixed rather than the past. No collision occurs when the moving ball is at $\vec{x}_2$ since we took the other ball out. We observe that those billiard balls which move straight from $\vec{x}_2$ to $\vec{x}_3$ were not at position $\vec{x}_1$ before they arrived at position $\vec{x}_2$. To be more precise, they were not at $\vec{x}_1$ before they arrive at position $\vec{x}_2$ such that the time interval between these two positions is $t_2 - t_1$. So, the collision is a backward cause of the billiard ball *a* being at position $\vec{x}_1$ at time t_1.

A symmetry problem thus arises: backward and forward interpretation are equally defensible options without the Humean convention. There is no asymmetry between the two interpretations if we remain completely open to the possibility of backward causation. This has unfavourable consequences for Woodward's interventionist account. Let D be a binary variable, saying that billiard ball a is at position $\vec{x}_1$ at time t_1 and has the corresponding velocity of the actual setting. C means that there is a collision between the billiard balls when a is at position $\vec{x}_2$ with b. Finally, E is a binary variable saying that a is at position $\vec{x}_3$ at time t_3 with the corresponding velocity of the actual setting. The interventionist test for causation between C and E is then not conclusive, even though the causal scenario is quite simple. We can make a case for the intervention on C changing the value of E, while D remains unchanged. But we can also make a case for the intervention on C changing the value of D, while E remains unaffected. Hence, it remains indeterminate whether or not an intervention on C changes the value of E. The conditional 'an intervention on C would change the value of E' lacks a determinate semantic value—without the Humean convention as a tiebreaker. As a result, the interventionist analysis does not tell us which event causes which other event for the simple causal scenario in question.

How to break the symmetry between forward and backward interpretation? Reichenbach (1956) made a simple suggestion:

> No wonder that acts of intervention change only the future, and do not change the past; the term 'intervention' is defined by the condition that the past be unchanged. The statement that acts of intervention cannot change the past is a trivial tautology. (p. 45)

Reichenbach thinks it is simply conceptually true that interventions do not change the past. This excludes the possibility of backward causation, given an interventionist account of causation. The benefit of adopting Reichenbach's suggestion is that it renders the interventionist test for causation determinate—at least for causal scenarios for which the laws of motion are known. Put differently, we eliminate an unacceptable indeterminacy from the semantics of interventionist conditionals by understanding the notion of intervention in such a manner that interventions do not change the past. If we understand interventions in this way and view causation from the perspective of an interventionist analysis, we resume a weak form of the

Humean convention about the direction of causation: an effect never precedes its cause. This weaker form of the Humean convention does not exclude cases of simultaneous causation, as we have seen in the previous chapter.

Some readers will have noticed that the general problem leading to the apparent indeterminacy of interventionist conditionals is due to the time symmetry of certain physical theories. Classical mechanics, classical collision mechanics, and some other theories are time-symmetric. That is, the laws of these theories remain valid if we change the direction of time. Such a change may be understood as follows: time point t is earlier than t' on the reversed direction of time iff t' is earlier than t on the customary direction. The variables t and t' stand for time points taken as primitive or real numbers conceived as representing time points. Russell (1913) took the time symmetry of fundamental theories in physics to be a reason for abandoning the notion of cause from science altogether. While Russell himself gave up on this radical position later on, he had a number of followers in the 20-the century and beyond.[3]

Unlike classical mechanics, thermodynamics is not time-symmetric. This fact is exploited in Reichenbach's proposal for defining the direction of time (Reichenbach 1956, Part III, IV). The presented argument concerning the direction of interventions is independent of whether or not we seek to define the direction of time. The argument goes through, even if we take the direction of time for granted. When developing his argument about the direction of interventions, Reichenbach himself wanted to show that interventions cannot define the direction of time. We have used the argument to defend a weak form of the Humean convention about the direction of causation.

As is well known, the symmetry problem concerning forward and backward interpretation arises for counterfactual accounts of causation in the tradition of Lewis (1973a). Lewis himself was well aware of the problem. He thought that the apparent symmetry is broken by what he describes as *overdetermination of the past by the future* (Lewis 1979). This idea has a number of severe problems to be discussed in sections 7 and 8 below. Let us, however, first review another argument against backward causation from

[3] See Frisch (2014, Ch. 5) for a detailed discussion and criticism of Russell's scepticism about causation, including also Neo-Russellian arguments against causation.

the perspective of the interventionist account.

4 The Bilking Argument

In this section, we explain the bilking argument against backward causation. This means we show that the following propositions entail a contradiction:

(1) It is impossible to change the past.

(2) There are cases of backward causation.

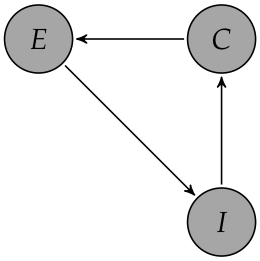
(3) Causation is understood in the sense of Woodward's interventionist account.

Suppose C is a direct cause of E. Further, suppose for contradiction, that E precedes C in our actual world. What happens if we intervene on C—relative to E—so that C does not occur? Whatever formal framework we choose to describe these changes—possible worlds or sets of value assignments—we have to conclude that E remains unaffected by the intervention on C because it's impossible to change the past. As we have seen in the previous chapter, the interventionist approach to causation must adhere to the principle that it is impossible to change the past. If changes of the past were to be admitted as possible consequences of hypothetical interventions, the interventionist account would become indeterminate as regards the direction of causation.

However, if we bring to bear other elements of the interventionist account of causation, we obtain the opposite result. Using the assumption that C is a direct cause of E, we can infer from (**M**) that setting the variable C to false—by an intervention I relative to E—changes the value of E to false such that this event does not occur anymore. Thus we have obtained a contradiction.

This proof by contradiction is intended to give the gist of the bilking argument. The original argument against backward causation by Black (1956) may be captured by the following causal model. Let I be the intervention variable by means of which we intervene on C. The intervention I sets the variable C to the value false if E occurs, and to the value true if E

does not occur. In more formal terms, we have the following causal graph and causal model *M*:



$$\boxed{\begin{array}{l} E = C \\ I = E \\ C = \neg I \end{array}}$$

Figure 72: Causal graph and causal model of the bilking argument

We can easily show that *M* is semantically inconsistent, that is, there is no value assignment to the variables such that all structural equations are true. Suppose *E* is true. By $I = E$, *I* must be true as well then. Using $C = \neg I$, we can infer from this that *C* is false. *C* being false and $E = C$ imply that *E* is false. Contradiction. Now, let us assume that *E* is false. By the same line of reasoning, we can show that *E* is true then. Again, we have obtained a contradiction. Since *E* must be either true or false, we have thus shown that there is no assignment of values such that all structural equations of the causal model are true. Notice that *M* is a cyclic causal model. There is a directed path from *E* to *E*. Unlike acyclic causal models, a cyclic causal model may not have a solution (Halpern 2000).

Notice that the bilking intervention satisfies all formal requirements for an intervention, as defined by (**IV**). *I* certainly causes *C*. Also, *I* trivially acts as a switch on all the other variables that cause *C*. Further, any directed path from *I* to *E* goes through *C*. Finally, *I* is independent of any variable *Z* that causes *E* and that is on a directed path that does not go through *C*. This holds true simply because there is no such variable *Z*. Hence, conditions I1 to I4 of (**IV**) are satisfied for the bilking intervention we envision. It is therefore a proper intervention in the sense of Woodward's account of causation.

We must wonder whether an analogous problem arises for temporally forward-directed causal relations. Suppose *C* causes *E*, and *C* precedes *E*. Further, suppose we intervene on *C* such that it is set to occur iff *E* does not occur. In the causal model that describes such an intervention, there must be a directed path such that one edge of this path is backward-directed in time. Hence, the intervention envisioned requires backward causation. So we can blame the inconsistency arising from bilking interventions on the admission of backward causation. To put it more carefully, admitting backward and forward causation in one and the same causal model allows for the construction of inconsistent causal models.

5 Ways out of the Bilking Argument

We have shown that the bilking argument against backward causation goes through if we adopt Woodward's interventionist account of causation. This is not to say that any interventionist account rules out backward causation. Price (1996) has given the bilking argument careful consideration, and yet tried to find a way out of the apparent consequences of this argument. We outline his strategy, and point out two problems, one of which is acknowledged by Price himself.

Price (1996, Ch. 7) distinguishes between two semantic conventions for the evaluation of interventionist counterfactuals. One says that we hold the past fixed when evaluating whether or not an interventionist counterfactual is true. This convention rules out backward causation in an interventionist account of causation by the bilking argument. Alternatively, we may hold the past fixed, but only to the extent it is epistemically accessible to us when evaluating an interventionist counterfactual. This weaker semantic convention opens up a loophole for backward causation. The bilking argument can be blocked if certain epistemic constraints are imposed on which interventions are admissible. More concretely, if the past effect is not known to us at time t, we cannot intervene on the cause at time t such that the effect would disappear. Without such epistemic constraints on interventions, the bilking argument goes through, as Price (1996, p. 180) himself has emphasized.

The next step is to show that we encounter phenomena in quantum mechanics such that past events are, in principle, epistemically inaccessible and certain measurements may be construed as causally influencing the past. In such phenomena two particles are emitted from a source S and some property P of one particle is measured at a time after the emission. Quantum mechanics predicts that the outcome of this measurement determines the outcome of measuring P for the other particle. This prediction has been confirmed experimentally. The result is surprising since the two particles are spatially separated so that no signal, travelling at the speed of light, could transmit any information from one measurement to the other. It is therefore difficult to conceive of a causal process going from one measurement to the other. This type of phenomenon is referred to as *quantum entanglement*.

Yet, non-standard accounts of quantum mechanics, known as *hidden variables theories*, open up the possibility of a causal interpretation along the following lines. The first measurement of P causally determines hidden variables concerning the source S at the time of emission. Then the values of these hidden variables causally determine the outcome of the second measurement of P, which concerns the other particle. The first leg of this causal process is backward-directed in time, while the second leg is forward-directed. The details are complex, and we do not attempt to give any introduction into quantum mechanics here. To appreciate our critical note to follow, it suffices to bear in mind that epistemic inaccessibility of the values of hidden variables—at the time when these variables assume respective values—is crucial for a coherent interpretation of backward causation in quantum mechanics. In the present example, it is certain properties of the source S—at the time of emitting two particles—which are captured by hidden variables. In Figure 73, M and M' stand for the measurements of property P for the two particles respectively, and S for the source from which the two particles are emitted:

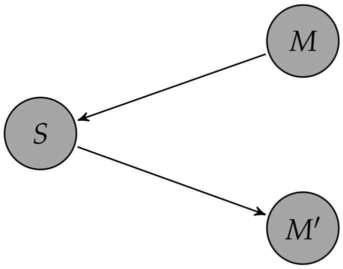


Figure 73: Backward causation in a scenario of quantum entanglement

The adumbrated account of backward causation severely limits the type of empirical evidence that can possibly be obtained for instances of backward causation. For this to be seen, notice that when testing for causation between X and Y by an intervention on X, we think that some changes in X

are accompanied with changes in Y. Now, if the past effect Y (the values of hidden variables concerning the source S) is inaccessible to us without measuring X (property P of one of the emitted particles), it is strictly impossible for us to observe any changes in Y due to an intervention on X. This means that we cannot observe any actual changes of the values of hidden variables concerning the source S, not even by using highly sophisticated measurement devices. Price is well aware of this problem, and emphasizes that arguments in favour of backward causation have to rely on holistic principles of theory choice, such as simplicity, elegance, and symmetry (1996, p. 180). Evidence for backward causation cannot have the form that changes in the presumed cause make us observe changes in the putative effect, not even indirectly.

More specifically, Price (1996, Ch. 9) argues that a combination of hidden variables and backward causation allows us to hold on to certain principles of locality when describing physical reality in quantum mechanics. Applied to the phenomenon of quantum entanglement, locality means that the outcomes of the two measurements may be understood in terms of individual properties of the two particles and the source of their emission. By contrast, the entanglement of the two measurement results cannot be construed as caused by individual properties of the two particles if causation is always forward-directed in time, even if hidden variables are used. This result follows from Bell's theorem.[4]

But even if we accept extraordinarily holistic limitations on arguments in favour of backward causation, there remains one more problem to be addressed. The problem arises from the following question: which interventions are admissible in an interventionist account of causation? Woodward discusses this question at greater length, eventually suggesting a notion of intervention which is maximally liberal and not constrained by the limitations of human agents or the laws of physics:

[4] For a more detailed account of quantum entanglements in the context of backward causation, the reader is referred to Price (1996, Ch. 8 and 9), Dowe (1997), Hausman (1998, Ch. 12.6), and Friederich and Evans (2019). The phenomenon of quantum entanglement seems to constitute the most powerful case for an interpretation in terms of backward causation (Dowe 1996, p. 228). But there are other phenomena in physics for which an interpretation in terms of backward causation has been explored. Faye (2021, Sect. 4) gives an overview. For a more technical discussion of hidden variables theories, see, e.g., Appendix F in Galindo and Pascual (1990).

> Commitment to a manipulability theory leads unavoidably to use of counterfactuals concerning what would happen under conditions that may involve violations of physical law. The reason for this is simply that any plausible version of a manipulability theory must rely on something like the notion of an intervention, and it may be that, for some causal claims, there are no physically possible processes that are sufficiently fine-grained or surgical to qualify as interventions. (Woodward 2003, p. 132)

In line with this liberalized understanding of interventions, Andreas and Casini (2019) made a more concrete case for interventions which violate laws of physics.

The need for physically impossible interventions leads to the following problem for an interventionist approach to backward causation: why should we refrain from hypothetical interventions on hidden variables? Such interventions are physically impossible, but so are other interventions which are needed in a comprehensive account of causation. Why should hypothetical interventions be constrained by what a human agent can know about hidden variables? Why is it not admissible to ask what happens if we had knowledge of hidden variables and used this knowledge to intervene on the presumed cause in a specific way?

From the perspective of Woodward's interventionist account, we have reason to admit bilking interventions. Woodward develops his interventionist account in such a manner that the human agent drops out in the end. Agency is not constitutive of causality (Woodward 2003, p. 126). This implies that interventionist conditionals are not constrained by epistemic limitations of human agents. Unlike Woodward, Menzies and Price (1993) retain the human agent in their interventionist approach to causation. At the same time, Price (2017, Sect. 4) points out that his interventionist account needs to include humanly impossible interventions inasmuch as Woodward's does. So, it remains an open problem on what grounds we should exclude interventions on hidden variables. Recall that one needs to exclude such interventions in order to block the bilking argument for causal relations among hidden variables. The envisioned loophole for backward causation closes if interventions on hidden variables are admitted.

One word on Dummett's subtle discussion of backward causation and the bilking argument is in order. Dummett (1964) aims to answer the following question: are there circumstances in which it can be rational to believe there is backward causation? His answer is yes, with one qualification: it must not be possible to know about the past effect without having intentions to bring it about or to prevent it. This implies that knowledge concerning the occurrence of the past effect is not available at the time of performing actions that are conceived to cause this effect. As just explained, Price (1996, chs. 7–9) has shown how hidden variable theories in quantum mechanics may be used to pursue this loophole for backward causation. At the same time, it should be noted that Dummett thinks that our notion of past rules out backward causation:

> The difference between past and future lies in this: that we think that, of any past event, it is in principle possible for me to know whether or not it took place independently of my present intentions; whereas, for many types of future event, we should admit that we are never going to be in a position to have such knowledge independently of our intentions. (Dummett 1964, p. 357)

Thereby, Dummett says that our conception of past violates a precondition for backward causation. Dummett's discussion of backward causation thus remains dialectical. He delivers arguments for both the proponent and the opponent to backward causation. On the one hand, Dummett (1964) agrees with the critic of backward causation, who holds that we can rule out this type of causation on conceptual grounds. On the other hand, he outlines a conceptual change of our notion of past which may open up the possibility of backward causation.

Gebharter et al. (2019, p. 132n) try to refute the bilking argument from an interventionist perspective within just a single paragraph. The case for deterministic backward causation goes as follows: if we prevent the occurrence of the cause—while the past effect already occurred—it remains true that the past effect occurred. This seems to contradict the interventionist account, according to which interventions on the presumed cause must affect the putative effect. Gebharter et al. (2019), by contrast, suggest inferring that 'another cause' of the past effect is (or was) present, which we have not been aware of and whose occurrence we did not prevent. In other words, we should conclude that we are in a scenario of overdetermination. This defence of backward causation is surprising. For it is well known that a counterfactual interventionist account of causation needs refinement when applied to scenarios of overdetermination. If we test for causation between X and Y in a scenario of overdetermination, overdetermining causes other than X must be set to non-actual values (see, e.g., Halpern and Pearl (2005)). Applied to the bilking scenario, this means that all overdetermining actual causes of the past effect must be intervened on when we test for causation between cause and effect. Otherwise, the interventionist account has no solution to the problem of overdetermination.

6 The Original Fork Theory

Reichenbach's fork theory aims to define the direction of causation using probabilistic and temporal information about event types. It was developed in his seminal *The Direction of Time* (1956). Reichenbach shows there that the direction of some causal relations can be derived from statistics, physics, and empirical observations. But there remain causal relations whose direction is determined by the Humean convention. We will show that this convention remains in place for a large set of causal relations. Backward causation is explicitly excluded.

The original fork theory by Reichenbach has been highly influential for approaches to causation which try to make do without the Humean convention or which deviate from this convention in one way or other (Dowe 1996, Papineau 1992, Price 1996, Spirtes et al. 1993). Some of these more recent approaches make room for backward causation (Dowe 1996, Price 1996). When developing his overdetermination thesis, Lewis (1979) does not explicitly draw on Reichenbach (1956), but this thesis turns out to be a deterministic variant of the original fork theory. For all these reasons, it is worth studying the original fork theory in greater detail.

The notion of a conjunctive fork is closely related to that of a probabilistic common cause. The idea is that the cause forms the head of a fork and the two causal relations form the tines, as depicted by Figure 74:

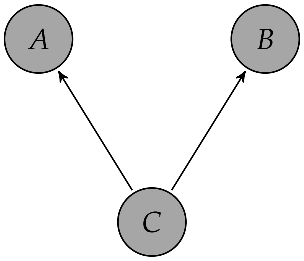


Figure 74: Fork with a common cause

The concept of conjunctive fork is defined as follows (Reichenbach 1956, p. 159):

Definition 31. Conjunctive fork
Let A, B, and C be types of events. A, B, and C form a conjunctive fork with the head event C—ACB is a conjunctive fork, for short—iff

(1) $P(A\&B \mid C) = P(A \mid C) \cdot P(B \mid C)$

(2) $P(A\&B \mid \overline{C}) = P(A \mid \overline{C}) \cdot P(B \mid \overline{C})$

(3) $P(A \mid C) > P(A \mid \overline{C})$

(4) $P(B \mid C) > P(B \mid \overline{C})$.

Conditions (1) and (2) say that A and B are probabilistically independent conditional upon C and $\overline{C}$. On assumption of C, the occurrence of A does not make the occurrence of B more likely or less likely. Nor does B make the occurrence of A more likely or less likely on this assumption. Likewise for the assumption that C does not occur, which is abbreviated by $\overline{C}$ (the complement of C). Conditions (1) and (2) are taken to imply that there cannot be a direct causal relation between A and B. Conditions (3) and (4) define some concept of probability raising. Condition (3) says that C raises the probability of A, or the absence of C lowers the probability of A, or both. In other words, A is more likely to occur when we know that C than it is when we know that not C. Condition (4) makes the corresponding assertion for B.

This concept of a conjunctive fork is insufficient to capture probabilistic common causes, though. For a causal chain of the following type satisfies conditions (1) to (4) as well:

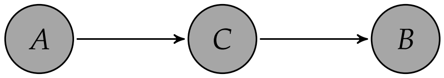


Figure 75: Causal chain

In such a chain, *A* and *B* are probabilistically independent conditional upon *C* and not *C*. And *C* raises the probability of both *A* and *B* in the sense of conditions (3) and (4). Reichenbach himself was well aware of this problem.[5]

Reichenbach's solution to this problem is twofold. First, he defines a probabilistic notion of *between*. While this notion is defined without reference to any causal notions, it has an intended causal meaning: *C* is between *A* and *B* iff there is a direct causal relation between *C* and *A* and one between *C* and *B*. In other words, *C* is between *A* and *B* iff there is an edge between *A* and *C* and an edge between *C* and *B* in the causal graph.

Second, Reichenbach takes recourse to temporal relations among events. More specifically, he makes the following two assumptions. Common causes have effects which approximately coincide in the sense that they occur together in a small spatiotemporal region. Likewise, common effects have causes which approximately coincide. He emphasizes that some principle of *local comparability of time order* is needed in his account of the direction of causation (p. 194). This is not to say that all effects of a common

[5] See Reichenbach (1956, p. 189), where the problem is explicitly acknowledged. Why did Reichenbach define the notion of conjunctive fork by conditions (1) to (4) in the first place, given he knew that causal chains satisfy these conditions as well? When discussing the probabilistic properties of common causes and common effects, Reichenbach makes the assumption explicit that the tine events *A* and *B* are simultaneous: 'The statistical relationships which two *simultaneous events* have, on the one hand, to a common cause, and, on the other hand, to a common effect, can be represented by the schema of Figure 23.' (p. 158, emphasis added) Figure 23 in Reichenbach (1956) shows two simultaneous events with a common cause and a common effect. Reichenbach then goes on to focus on the common cause and the relations to its effects. On the next page, the concept of a conjunctive fork is defined, albeit not in a maximally explicit manner. One reading of Reichenbach's terminology is that the simultaneity of the tine events within a fork remains an implicit assumption when the notion of a conjunctive fork is defined. Thereby, causal chains are excluded from qualifying as conjunctive fork. But there are also passages where the notion of conjunctive fork is used in a wider sense, which may well include causal chains: 'For this reason, we shall say that relations (5)–(8) [i.e, (1) to (4) in our definition] define a *conjunctive fork*, that is, a fork which makes the conjunction of two events *A* and *B* more frequent than it would be for independent events.' (Reichenbach 1956, p. 159)

cause coincide. Such effects may well be spatially and temporally separated from one another. Even then, Reichenbach argues, we can recognize intermediate events between the common cause and its effects, respectively, which do coincide. To be more precise, if A and B are effects of a common cause which do not coincide, then there are intermediate events A' and B' which do. A' is intermediate between C and A, and B' between C and B.[6]

So far, we have three types of information in order to determine the causal graph for a set of event types: (i) the probabilistic concept of between, (ii) the probabilistic concept of conjunctive fork, (iii) and temporal information about local coincidences. These three ingredients define what Reichenbach calls a *net with a lineal order* (p. 197). To understand what he means by such a net, let us begin with a simple example (cf. Figure 28 on p. 181 in Reichenbach (1956)):

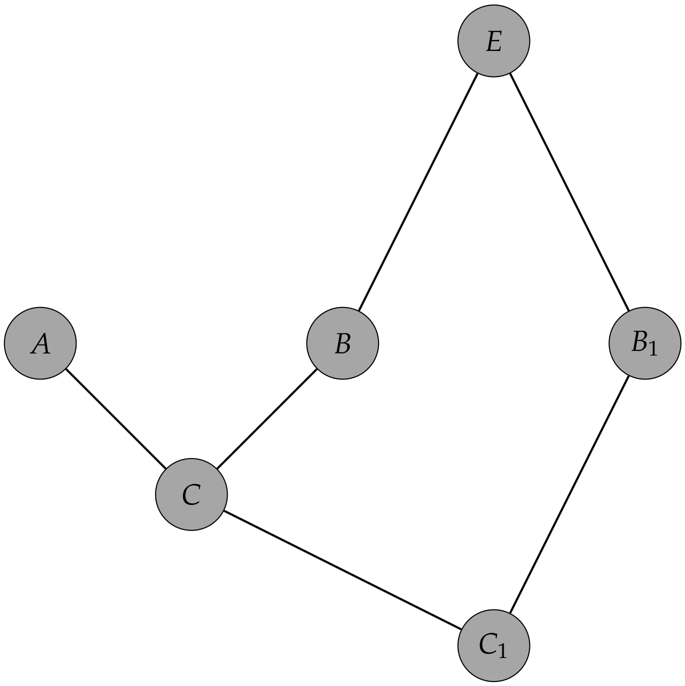


Figure 76: Combination of an open fork with a closed fork

As a first approximation, a lineally ordered net may be described as an

[6] See Reichenbach (1956, Ch. 22) for details.

undirected causal graph. The edges stand for direct causal relations, but we do not know in which direction causation goes. Take Figure 76 as an example, which combines an open fork with a closed fork. How do we assign a direction of causation to the edges of the net?

Based on broadly empirical observations and assuming our customary understanding of time direction, Reichenbach claims that we have conjunctive forks open to the future, but no such forks which are open to the past. A fork ACB is open to the future iff the two tine events A and B are temporally later than C, but there is no event E such that AEB form a conjunctive fork and E is temporally later than A and B. This observation suggests the following definition of the direction of time: 'In a conjunctive fork ACB which is open to one side, C is earlier than A or B' (Reichenbach 1956, p. 162). This definition enables us to assign a direction of time to the edges between C and A, and C and B in the net of Figure 76. Using this assignment, a direction of causation is assigned to the undirected edges between C and A, and C and B too. Thus we obtain the net of Figure 77:

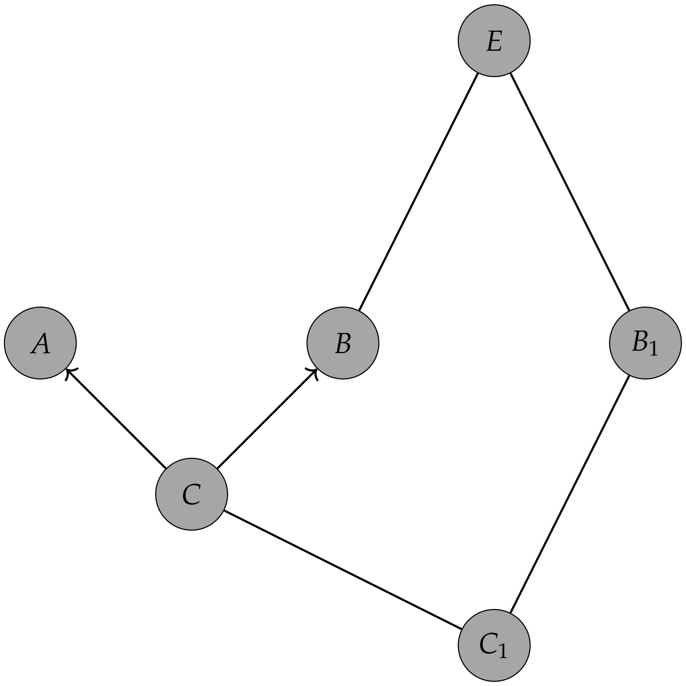


Figure 77: Combination of an open fork with a closed fork, where a direction is assigned to the edges in the open fork.

It is not entirely clear, which assignment comes first—that of time or that of causation. In any case, the direction of time and the direction of causation are aligned with one another such that the Humean convention is satisfied.

This definition of time in terms of open conjunctive forks complements the statistical definition of time direction, according to which the arrow of time is aligned with the direction of growth of entropy. That is, if the entropy of a system changes from a lower state to a higher, the lower state precedes the higher. Entropy grows in the universe and the various subsystems thereof. Let's call this definition *entropy time*. Furthermore, Reichenbach tries to show—for statistically irreversible processes—that the direction of causation is aligned with the direction of entropy time (1956, Part III).

We are almost done. The last element in Reichenbach's account of the direction of causation is the following principle: once we have directed a single edge in the net of events (defined by relations of *between*, conjunctive forks, and approximate coincidences), all the other edges are directed as well. This is what Reichenbach means by the notion of a *lineally ordered net*:

> the causal net constructed statistically has acquired the same properties as the causal net of classical mechanics: it is ordered as a whole. The net thus possesses a *lineal order*. This means that, if a time direction is assigned to one causal line, a direction results for every line. (Reichenbach (1956, p. 197), emphasis added)

Why can we determine the direction of all edges on the basis of assigning a temporal direction to a single edge? So far, we have established the following claims: (i) the direction of causation is aligned with the direction of entropy time. (ii) There are no conjunctive forks open to the past in the sense of entropy time. (iii) Entropy time is understood such that entropy grows over time: the states with higher entropy are later than states with lower entropy. (i), (ii), and (iii) imply that all open conjunctive forks ACB of the net are such that the head event C is a cause of A and B. Thereby, we have directed all edges of all open forks.

However, not all edges of the net form a tine of an open fork, as is exemplified by the net of Figure 77. Why does the directedness of open forks allow us to assign a direction to the other edges in the net? To answer this question, it is crucial to note that a lineally ordered net is not merely an undirected graph in which the edges stand for causal relations. Such a net rather contains information about temporal relations among events. More specifically, a given lineally ordered net allows for only two interpretations of time order. These interpretations may be represented by two directed acyclic graphs G_1 and G_2. G_1 and G_2 have the following properties. First, if there is a directed edge from A to B in G_1, then there is a directed edge from B to A in G_2. And vice versa. No graph contains any cycles. In particular, there are no nodes A and B such that there is a directed edge from A to B and vice versa in one and the same graph. Graph-theoretically, a net with a lineal order is a pair of two directed graphs which have said properties.

Reichenbach represents the two temporal interpretations of a lineally ordered net graphically, but not graph-theoretically. Suppose there is an edge between A and B such that A is closer to the top of the page (on which the net is depicted) than B. Then A is temporally later than B or temporally earlier. Moreover, all pairs (AB) and (CB) of edges in the net have the following property: if A is higher up on the page than B and C higher up than D, then either A is later than B and C is later than D, or B is later than A and D is later than C. It is then excluded, for example, that A is later than B and D later than C. This is, we suggest, what Reichenbach means by saying that the net is ordered as a whole: once one edge is assigned a direction of time and causation, all the other edges are directed as well.

In sum, causation and time either go from the bottom to the top of the page, or vice versa. Put more technically, the diagram of a lineally ordered net has the following property: the direction of time is either parallel to the direction from the bottom to the top of the page or antiparallel to this direction. Likewise for the direction of causation. The main reason for this interpretation is as follows. Suppose we understand the notion of net with lineal order graph-theoretically such that the net is given by a single undirected graph. That is, all edges of the net are undirected, and we have no information about temporal relations among events. Then it is simply wrong to say that assigning a direction to one edge determines the direction of all the other edges. This is even wrong if we assume that all edges must point in the same direction after the assignment. For the concept of sameness of direction is not well defined if the net is understood as a single undirected graph.

Our interpretation is furthermore confirmed by the fact that all diagrams of directed causal graphs in Reichenbach (1956) conform to the convention that the events higher up in the diagram are temporally later than the lower events. Notably, there are no diagrams of causal graphs with a directed edge between *A* and *B* such that the edge is horizontal. Information about temporal relations among events can be obtained via relations of approximate coincidence, as indicated above. Suppose there is an edge between *C* and *A*, *C* and *B*, and *C* and *D*. Further, *A* and *B* coincide or there are intermediate events—intermediate between *C* and *A* and intermediate between *C* and *B*—which coincide, or both. However, *A* and *D* do not coincide, and there are no intermediate events—intermediate between *C* and *A* and intermediate between *C* and *D*—which do. Hence, either *A* and *B* are higher up than *C* and *C* is higher up than *D*, or the other way around (*A* and *B* are lower than *C* and *C* is lower than *D*). These considerations can be extended to the whole net such that the direction from the bottom to the top of the page represents the direction of time or the opposite direction of time.

We have thus answered the question of how those edges are directed which are not tines of an open fork. These edges must be aligned with the edges which are tines of open forks. Given a lineally ordered net, alignment can only mean that the edges not in open forks have the same temporal direction as the edges in open forks. For example, this convention of alignment enables us to assign a direction of causation to all edges in the net of Figure 77. First, by the above definition of the direction of time in terms of open forks, we know that *C* is prior to *A* and *B*. Then, by the convention that edges in open and closed forks need to be aligned, we can assign a direction of time to the other edges. The direction of time immediately gives us the direction of causation. The two assignments are represented by the following figure:

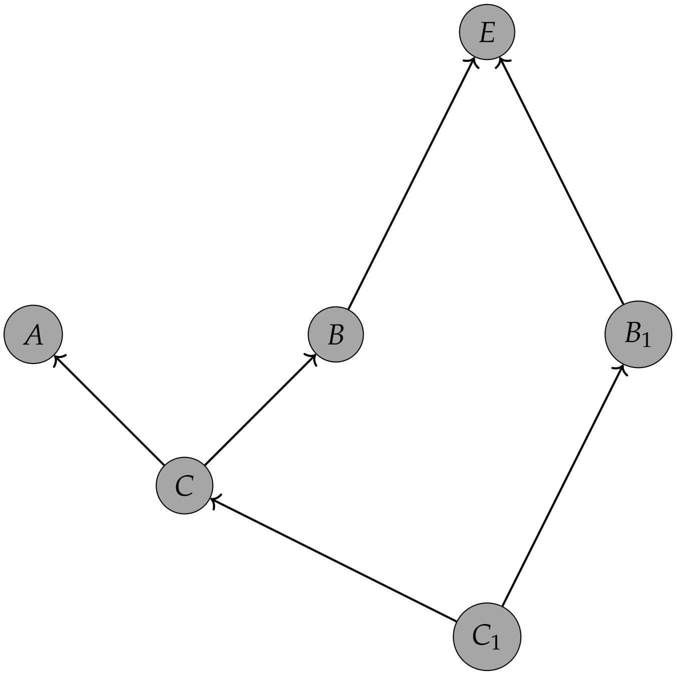


Figure 78: Combination of an open fork with a closed fork, where a direction is assigned to all edges.

Note that Reichenbach's alignment convention makes use of the Humean convention to assign a direction of causation to edges which are not in an open fork. Reichenbach himself says that we confer a direction of causation on reversible processes by the convention that this direction is aligned with the causal direction of irreversible processes and open forks (p. 156). Moreover, he holds that our world is such that causation is always forward-directed in time or always backward-directed in time.[7] Once entropy time and fork time are established in the above manner, we obtain that causation is always forward-directed in time. There is no loophole for backward causation. Even simultaneous causation is difficult to conceive.

The problem of how to assign a direction to edges which are not in open forks turns out trouble for more recent accounts of causation. A case in point is Lewis's counterfactual account to be discussed in the next two sec-

[7] See Reichenbach (1956, pp. 35–9, p. 153n). This is very much in line with our claim that the notion of lineal order is applicable only if we do not have both forward and backward causation in one and the same causal scenario.

tions.

7 The Counterfactual Approach

Using his counterfactual approach to causation, Lewis (1973a, 1979) tried to derive the direction of causation from the semantics of counterfactuals themselves. If successful, this derivation achieves two objectives. First, Lewis can explain why causation is almost always, if not outright, forward-directed in time. Second, exceptions from forward-directed causation become conceptually possible. The project is even more ambitious and heroic than Hume's is. The objective is to give a reductive analysis of causation without relying on the asymmetry of time.

The notion of backtracking counterfactual plays a paramount role in Lewis's reductive analysis. What is a backtracking counterfactual? There are at least two readings: a causal and a temporal one. On the causal reading, a backtracking counterfactual goes, by definition, against the direction of causation. On the temporal reading, such a counterfactual goes, by definition, against the direction of time. It's far from easy to figure out which of the two readings is correct, specifically since the familiar causal relations are all forward-directed in time. Let C be a cause of E such that C precedes E. Then the counterfactual 'had E not occurred, C would not have occurred either' is backtracking in both the causal and the temporal sense. 'Had C not occurred, E would not have occurred either' is non-backtracking in both the causal and the temporal sense of backtracking.

We tried to make a case for the temporal reading of backtracking, but have to admit that there are also passages in Lewis (1979) which support the causal reading. The problem is that Lewis (1979) often uses both causal and temporal notions in order to explain backtracking. If backtracking concerns, by definition, the direction of causation and not the direction of time, it would have been helpful had Lewis explained the distinction without references to the temporal relation between cause and effect. If backtracking concerns, by definition, the direction of time and not the direction of causation, it would have been helpful had Lewis explained the distinction without references to causal relations. After all, it is a matter of convention how we understand the notion of a backtracking counterfactual. To minimize confusion, we suggest distinguishing more clearly between causally and temporally backtracking counterfactuals.

Lewis tried to show that causally backtracking counterfactuals are always false at our world, given his proposed semantics of counterfactuals and a suitable account of similarity among worlds. To understand his reasoning, let us review very briefly the basic elements of the semantics of variably strict conditionals by Lewis (1973b). Suppose A and B are false at world w. Let's call a possible world where α is true an *α-world*. A counterfactual $A \mathrel{\Box\kern-1.5pt\rightarrow} B$ is true at a possible world w just in case some $A \wedge B$-world is *more similar* to w than any $A \wedge \neg B$-world, if there are any A-worlds. If there are no A-worlds, the counterfactual $A \mathrel{\Box\kern-1.5pt\rightarrow} B$ is vacuously true. If the set of possible worlds is finite, we can say that $A \mathrel{\Box\kern-1.5pt\rightarrow} B$ is true at w iff all A-worlds which are most similar to w are B-worlds, or there are no A-worlds. This semantics is based on a similarity ordering among possible worlds. Lewis (1979, p. 472) suggests imposing the following constraints on this ordering:

(1) It is of the first importance to avoid big, widespread, diverse violations of law.

(2) It is of the second importance to maximize the spatiotemporal region throughout which perfect match of particular fact prevails.

(3) It is of the third importance to avoid even small, localized, simple violations of law.

(4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly.

These constraints are motivated by several objectives. First, agreement with our counterfactual judgements when the formal semantics is applied to counterfactuals in natural language. Second, eliminating some cases of vagueness concerning the meaning of counterfactuals in natural language. Third, a plausible theory of causation when the similarity ordering is used to complement Lewis's (1973a) analysis of causation.

So far, everything is symmetric. We have no reason to think that counterfactuals going against the direction of time and causation come out false on Lewis's semantics. The symmetry is broken by the overdetermination thesis: the past is overdetermined by the future—at least in our world and perhaps with a few exceptions. The exceptions are supposed to make room for backward causation, as we shall see shortly. Overdetermination of the past may well be understood in analogy to familiar cases of causal overdetermination, which are forward-directed. In these familiar cases, several events $C_1, \ldots, C_n$ are individually sufficient to bring about a certain effect E. Recall some stock example of overdetermination: several soldiers shoot a prisoner, and each bullet is fatal. In these familiar scenarios of overdetermination, there are several events C_i such that C_i determines that a future event E occurs, given the laws governing our world. In cases of overdetermination of the past, by contrast, we have several events E_i such that each E_i determines the occurrence of a prior event C, given the laws of our world. This type of determination must not be understood causally, but rather nomologically. Using the convention from Reichenbach (1956) that the direction of time goes from the bottom to the top of the page, we can graphically depict the two types of overdetermination by the following figures:

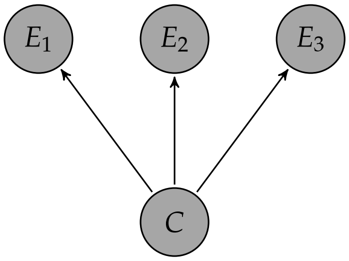


Figure 79: Overdetermination of the past

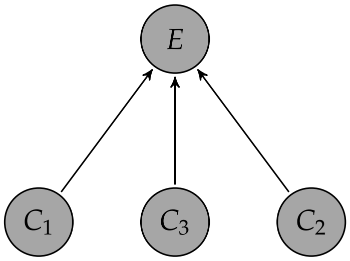


Figure 80: Overdetermination of an effect E

Figure 79 is a schematic representation of the temporal and causal relations of overdetermination of the past. It is assumed that the events E_1, E_2, and E_3, individually, allow us to infer back to the past cause C. Figure 80, by contrast, represents the temporal and causal relations in the familiar cases of overdetermination, given that C_1, C_2, and C_3 are individually sufficient to bring about E. The figures assume that it is single events—as opposed to conjunctions of events—which determine the occurrence of another event. This is in line with the paradigmatic examples adduced by Lewis for the overdetermination thesis. The assumption suggests itself if we read Lewis's approach to the direction of causation as fork theory. Below we will also look at conjunctive scenarios in the context of the overdetermination thesis.[8]

Now, Lewis thinks that overdetermination of the past is prevalent, if not omnipresent, while overdetermination of a future event is rare. He motivates this thesis of overdetermination with reference to the physics of wave propagation. The source of a wave is given by an event within a relatively small and well-confined spatiotemporal area. Take a stone dropping into a pond or emission of light by a fire. The waves originate from the source and spread themselves through a relatively large area in spacetime. There are numerous events later than the original creation of the wave from which we can infer that there is a certain source with specific properties, such as wave length and frequency.

In fact, the overdetermination thesis, if correct, allows us to explain why causation is predominantly forward-directed in time. Suppose C is a cause of several events E_i $(1 \leq i \leq n)$, which are temporally later. Since causation is asymmetric by assumption, none of the events E_i $(1 \leq i \leq n)$ is a cause of C. Further, each E_i determines—together with the laws—that C occurred. The occurrence of C is not overdetermined by past events, though. None of the effects E_i $(1 \leq i \leq n)$ is overdetermined. Let us now apply the simple counterfactual test for causation to this scenario, setting aside chains of counterfactual dependence for simplicity. Then $\neg C \; \Box\!\!\rightarrow \neg E_i$ must come out true for all i $(1 \leq i \leq n)$. This is quite plausible since avoidance of miracles is crucial when it comes to assessing similarity among possible worlds. Since C is not overdetermined by a past event, roughly just one miracle is needed to deviate from the actual world to a world where C is false. In such a possible world, none of the events E_i $(1 \leq i \leq n)$ should occur if we want to avoid further miracles.

What would have happened if some event E_i $(1 \leq i \leq n)$ had not occurred? As regards the past of E_i we have at least three principal options. First, assuming a miracle shortly before E_i occurs such that C still occurs in the corresponding possible world. Let's call this set of possible worlds W_1. Second, assuming a miracle shortly before C such that C does not occur, while the events E_j other than E_i still occur. Here, the events E_j $(1 \leq j \leq n, j \neq i)$ occur miraculously since they are uncaused. Let's call the set of these pos-

[8] In reading Lewis's (1979) proposal as fork theory, we follow Price (1996, Ch. 6) and Papineau (1992).

sible worlds W_2. The third option is to assume a miracle shortly before C such that neither C nor any of the events E_i $(1 \leq i \leq n)$ occurs. Let's refer to the corresponding set of possible worlds by W_3. Clearly, the members of W_1 are most similar to the actual world, provided our assessment of similarity is confined to the set $\{C, E_1, \ldots, E_n\}$ of events. The members of W_2 involve more miracles than the members of W_1. Also, the worlds of W_2 agree with the actual world completely as regards particular fact. The worlds in W_3 do not differ from those in W_1 as regards miracles, but the worlds in W_1 agree better with the actual world as regards particular fact than the worlds in W_3 do. This gives us reason to deny that the counterfactual $\neg E_i \mathrel{\Box\!\!\rightarrow} \neg C$ is true, as it should be. After all, the counterfactual is backtracking in both the temporal and the causal sense.

In sum, the fork structure of causal relations leads to an asymmetry of miracles. If we counterfactually assume that the head event of the fork had not occurred, we should hold that none of the tine events would have occurred either. For a number of miracles are needed if the tine events were to occur without the head event. Hence, we have reason to think that the counterfactual 'had the head event not occurred, the tine events would not have occurred either' is true. By contrast, if we counterfactually assume that a tine event had not occurred, we should hold that the head event would still have occurred. For only a small miracle is needed in order to maintain complete agreement about particular fact—as regards the past of the tine event—between the actual world and the possible worlds of the counterfactual assumption that the tine event had not occurred. If we can achieve complete match of particular fact as regards a large temporal interval by just one small miracle, then this is 'better' than complete avoidance of miracles at the price of large disagreement about particular fact. Hence, we have reason to think that the counterfactual 'had a tine event not occurred, the head event would not have occurred either' is false. Lewis's counterfactual account of causation tells us that the head event is a cause of the tine events, but not vice versa.

For Lewis's overall account of causation to work, it is crucial that small miracles are tolerable, but big and widespread miracles are not. The latter type of miracles are referred to as *big, widespread, diverse violations of law* (Lewis 1979, p. 472). Suppose Lewis were to say that, for the assessment of similarity, it is of the first importance to avoid violations of law outright. Then a possible world with a small miracle would always be less similar to the actual world than any possible world without miracles at all. A large number of (causally and temporally) backtracking counterfactuals would come out true then. There would be no asymmetry of causation.

Lewis's theory of causation succeeds in making room for the conceptual possibility of backward causation, at least on a charitable reading. Overdetermination of the past by the future may be violated locally. If so, we seem to have a case of backward causation. Of course, the question arises of how to distinguish backward causation from the familiar cases of forward-directed overdetermination. But there is at least some prospect of a coherent understanding of forward-directed causation which allows for exceptions in the form of backward causation.

8 Problems of the Counterfactual Approach

If causation has a fork structure, we must wonder whether forks are omnipresent. Is there a unique fork for every cause such that the head event causes the tine events, while every single tine event determines that the head event occurs—given the laws of our world? Put in simpler terms, is every cause the head of a fork? Lewis (1979) himself does not discuss this question. Papineau (1992, p. 239), however, maintains we should understand Lewis's fork theory in precisely this way:

> given *any* event *C*, in one direction in time there will be many different sequences each of a type which is generally found with *C*, while in the other direction in time there will only be one such sequence. And then we can say that the former sequences of events are the effects of *C*, and the latter sequences its causes. (Papineau (1992, p. 239), our emphasis)

On this reading forks are omnipresent: any cause is the head of some fork. We show that Papineau's reading is correct. Every causal relation must be embedded in a corresponding fork if we accept Lewis's counterfactual account of causation, the formal semantics of counterfactuals, and the underlying similarity ordering of possible worlds. Suppose, for contradiction, we have a causal structure depicted by Figure 81:

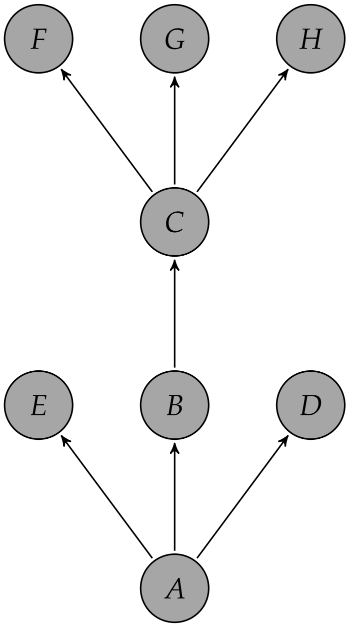


Figure 81: Two forks connected by single directed edge

By assumption, B causes C, and the causal relation between B and C is not embedded in any fork. Let us apply the counterfactual test to this causal relation. What would have happened if B had not occurred? To answer this question, we must look at the $\neg B$-worlds which are most similar to the actual world.[9] Given the causal structure of the scenario, it's obvious which of the $\neg B$-worlds are to come out as closest to the actual world. To get the causal structure right, the $\neg B$-worlds closest to the actual world are such that a small miracle occurs shortly before B but no miracle occurs shortly thereafter. The problem is that Lewis's constraints on the similarity ordering—explained in the above section—do not suffice to obtain this result. For this to be seen, consider a possible world w_1 where just one miracle occurs shortly before the absence $\neg B$ and the future unfolds according to the laws. To be precise, there is a miracle in w_1 shortly before the time point at which we expect B to occur from the perspective of the

[9] We tacitly assume that only a finite number of possible worlds is considered, which simplifies the semantics of counterfactuals as indicated in the above section. This assumption is justified since we consider only a finite number of events. Even if the assumption is not justified, the following considerations are easy to translate to the proper semantics of counterfactuals for an infinite number of possible worlds.

actual world. Further, consider a possible world w_2 in which a miracle occurs shortly before the absence $\neg B$ and shortly thereafter so that C occurs after all. The two possible worlds w_1 and w_2 compete for similarity. Which possible world wins the competition?

Let w_0 be the actual world. Clearly, neither w_1 nor w_2 exhibits *big, widespread, diverse violations of law*. So the first criterion of similarity is satisfied by w_1 and w_2 equally well. Further, there is almost complete and perfect *match of particular fact* between w_2 and w_0. But there is less perfect match of particular fact between w_1 and w_0. This is confirmed by how Lewis (1979) himself applies the similarity criteria to a concrete example, as we will see shortly. The third criterion of similarity—to avoid *even small, localized, simple violations of law*—is satisfied by w_2 to a lesser extent than w_1 since there is just one miracle in w_1 but two in w_2. The fourth criterion—*to secure approximate similarity of particular fact*—is not really applicable since our scenario is discrete. Even if it was applicable, this criterion would not be decisive since the two worlds do not satisfy all of the higher criteria equally well.

The surprising result is that the world with two miracles—one shortly before and another shortly after the absence $\neg B$—wins the competition. w_2 is more similar to the actual world than w_1 since it exhibits a larger spatiotemporal region with perfect match of particular fact, while neither world exhibits big, widespread, diverse violations of law. This, however, implies that the counterfactual test fails to capture that B causes C, and so it also fails to capture that B is a cause of F, G, and H. Notice that even a tie between w_1 and w_2 would be a severe problem for Lewis. To capture that B causes C, Lewis has to show that w_1 is more similar than w_2. Thus we have received a contradiction from the assumption that not all causal relations are embedded in a fork.

It is important to note that our assessment of similarity concerning w_1 and w_2 is confirmed by how Lewis (1979) himself applies the similarity constraints to a concrete example. As is well known, Lewis (1979) tried to show that the counterfactual 'if Nixon had pressed the button there would have been a nuclear holocaust' comes out true, given his constraints on the similarity order and a few auxiliary assumptions about the button and its connection to nuclear weapons. Lewis's account of similarity is in part motivated by the objective to deliver the intuitive result for this counterfactual.[10] Among the possible worlds in which Nixon does press the button, there are ones in which a small miracle occurs before Nixon presses the button and some miracles occur thereafter such that no holocaust occurs. Why are such possible worlds less similar to the actual world than those in which no miracle occurs after Nixon pressed the button? To answer this question, Lewis points out that Nixon's pressing the button leaves several traces (Lewis 1979, p. 469–71). In other words, Nixon's pressing the button is a head event within a fork. To formally capture the causal efficacy of Nixon's pressing the button—in a hypothetical scenario where he does so—it is therefore important that the hypothetical cause is embedded in a fork. These considerations apply to actual causes analogously.

Let us return to Figure 81, and the causal relation between B and C in the corresponding causal scenario. We have seen that Lewis's overall package of similarity constraints, semantics of counterfactuals and a counterfactual approach to causation fails to capture this relation as causal since it is not embedded in any fork. We must wonder whether the problem could be solved by making the constraints on the similarity order more precise. In the context of the problem to be solved, we see only two principal options for this to be done:

(1) Declaring that a possible world with at least one miracle less than another is always more similar to the actual world than the latter.

(2) Declaring that it is of first importance that the number miracles remains below a certain threshold which is exactly specified.

Let us begin with the first option. If adopted, it could be shown that w_1—the $\neg B$-world with just a miracle shortly before $\neg B$—is more similar to the actual world than w_2 in which two miracles occur. However, possible $\neg B$-worlds without any miracles would then count as more similar to the actual world than w_1. In such worlds, not only the future of $\neg B$ deviates from the actual world, but also larger parts of the past of $\neg B$. (By future of $\neg B$ we mean the time point or time period in which event B is absent compared to the actual world.) As a consequence, option (1) leads to the problem that backtracking is omnipresent. Many backtracking counterfactuals—in both the temporal and the causal sense—would come out true on this option.

[10] Lewis (1979) responds here to a critical note by Fine (1975).

Notably, Lewis (1979, p. 469) himself considers the option of understanding similarity in such a manner that miracles are to be avoided at all costs. He rejects this option because it leads to unlimited backtracking.

If we chose option (2), we would have to declare that a possible world with more than one miracle is always less similar to the actual world than a possible world with just one miracle. With this ad-hoc convention, we could show that w_1 qualifies as more similar to the actual world than w_2. However, the convention leads to new problems at the level of counterfactuals. Suppose we are in a scenario of forward-directed overdetermination such that C and A are overdetermining causes of E. Since there is no backward causation in this scenario, we expect that the following counterfactuals come out true: 'had E not occurred, C would still have occurred' and 'had E not occurred, A would still have occurred'. However, it is easy to show that these counterfactuals come out false on the convention that at most one miracle is tolerable. Likewise, it can be shown that the following backtracking counterfactual comes out true on the convention: 'had E not occurred, C or A would not have occurred either'. The general problem is that—in order to block backtracking in a scenario of forward-directed overdetermination of event E—more than one miracle is needed—in the possible $\neg E$-worlds considered most similar to the actual world.

Thus we have arrived at the conclusion that every causal relation must be embedded in a fork on Lewis's overall account of causation. This, however, implies that every cause produces several effects, each of which in turn brings about several further effects, and so on. If this picture were correct, there would be unlimited exponential growth of events. Figure 82 is an attempt at a simple graphical explanation of why the number of concurrent events grows exponentially if every causal relation is embedded in a fork:

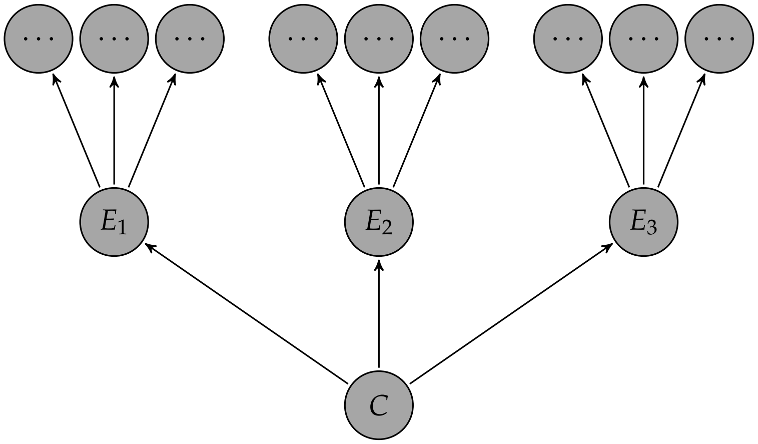


Figure 82: Exponential growth of concurrent events

Now, causation cannot be bound to imply exponential growth of concurrent effects. At least if we take our causal judgements concerning well established theories in physics seriously, the picture cannot be correct. A case in point is classical mechanics. Take the Moon orbiting around the Earth. We take it that gravitational forces between Moon and Earth are among the causal factors which determine the orbit—unless we follow Russell (1913) in denying causation to be a legitimate concept of physics. The problem for Lewis's account is that there is no growth of events whatsoever in the idealized model in which Earth and Moon are considered to be point particles.

Are less idealized systems of the system of Earth and Moon any better off in the sense that some growth of events becomes recognizable for them? Such systems consider Moon and Earth as spatially extended objects and take various kinds of friction into account. For example, they consider that the momentum of the Moon is slowed down by the tides which are caused by the Moon. Even for such less idealized systems, exponential growth of effects is very difficult to recognize. The mere phenomenon of tides does not involve any growth of events. Considerations of the growth of entropy may or may not lead us to recognizing some growth of events, but not for smaller temporal intervals.

More generally, it is worth noting that most, if not all, systems in classical mechanics do not show any growth of concurrent events. Think of a system of billiard balls modelled by classical collision mechanics. There is a canonical description of such a system in terms of the momentum and the position of each ball at a given time point. If we calculate the future development of the system using laws of classical collision mechanics, we do not obtain any growth of concurrent events. The number of events—given by the canonical description—rather stays constant. Similar considerations apply to systems of classical mechanics used as foundation in statistical physics and thermodynamics.

Could Lewis's account be defended by switching from classical to relativistic mechanics? We were not able to see a loophole opening up by this move, but have to admit that the concept of concurrent events becomes relativized then. In any case, if we take our causal judgements seriously—including our causal judgements for idealized systems in science and everyday life—a proper theory of causation must not end up with the conclusion that there is no causation in classical mechanics and many other models of causal processes.

Finally, there is a more general, cosmological problem with the conclusion that there is unlimited exponential growth of concurrent events, which is entirely independent of our causal judgements. This type of growth is not sustainable, given there is only a limited number of particles in the universe. Even under moderate assumptions about the parameters of forks in our world, the number of concurrent events would exceed the number of particles in our universe after just a few years of exponential growth. We have run a little simulation, the result of which was that 100 events produce $2.4 \cdot 10^{147}$ concurrent events after only seven years of causal efficacy. The details are not very interesting, and so we leave them out. The interested reader is encouraged to devise her own simulation.

It is also instructive to take a look at familiar cases of exponential growth in nature and society. A bacteria culture grows exponentially as long as enough nutrition is available. The coronavirus spread exponentially for a certain time in some countries in 2020 and 2021. Some economies showed exponential growth after World War II. The population of certain countries with high birth rates and improved medical care has grown exponentially in the past. However, in all of these cases, exponential growth comes to an end eventually. The growth rate of a bacteria culture will decline if not enough nutrition is available anymore. Periods of economic exponential growth are followed by economic crises. Even if no measures had been taken against the spread of the coronavirus, the number of new cases of

COVID-19 would not have grown exponentially forever. At a certain stage, there are not enough people left to sustain unlimited exponential growth of new COVID-19 cases. There is no unlimited exponential growth in nature and society. This type of growth only exists in our theoretical models. It is therefore, to say the least, extremely implausible to assume unlimited exponential growth of concurrent events in our physical world.

For the sake of transparency, we should make two assumptions explicit which have been made in this section. First, most effects have further effects. Second, forks do not massively overlap in such a manner that a given event is always the endpoint of several tines coming from different forks. Figure 83 is an attempt to graphically illustrate this kind of overlap. This figure depicts only two forks which overlap symmetrically rather than a larger number of forks which overlap in arbitrary ways. Our discussion, however, does not lose generality by considering just two interrelated forks.

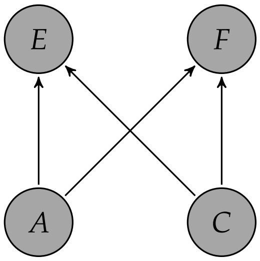


Figure 83: Overlap of two conjunctive scenarios

The first assumption is easy to justify. We can almost always think of causal scenarios in which a given effect has further causal consequences. Think of a rock which has been moved a few centimetres on the ground. As a consequence of this, certain air molecules collide with the rock which would not have done so if the rock had stayed at the original place. Certain light waves get absorbed or reflected by the rock which would have been absorbed or reflected in a different way by the ground. In the discussion of the Nixon example, Lewis (1979, 469n) himself assumes that even minor effects have further causal consequences.

What about the objection that forks may massively overlap so that no exponential growth follows from the overdetermination thesis? In Figure 83, the effects result from a conjunction or a disjunction of events. We can rule out the disjunctive case since then there would be as much overdetermination of the future as overdetermination of the past. So, *A* and *C* must be conjunctive factors of both *E* and *F*. It must also hold that we can infer *A* and *C* from *E* and the laws of the causal scenario. Likewise we must be able to infer *A* and *C* from *F* and the laws. Otherwise there would be no overdetermination of the past. If *E* and *F* were needed to infer *A* and *C*, there would be no such overdetermination.

The problem is that the latter conditions—saying that the conjunctive factors are inferable from a single effect—are rarely met by conjunctive scenarios in our world. Most of the time, at least some conjunctive factors are not inferable from the effect since a given effect may be brought about in various ways. Classical collisions are a striking example. Suppose *a* collides with *b* so that both *a* and *b* change their momentum because of the collision. We have at least two conjunctive causal factors: position and momentum of *a*, and position and momentum of *b* at a time prior to the collision. And we have at least two effects: position and momentum of *a*, and position and momentum of *b* at a time after the collision. Now, there is no way to infer the two conjunctive factors from a single effect using the laws of collision mechanics. The fact that *a* has a certain position and momentum at a certain time point after the collision may be brought about in lots of different ways. The actual conjunctive causal factors are just one possibility which happens to be actual. In order to infer conjunctive causal factors from effects, we need to know, at least, both effects. Position and momentum of both *a* and *b* at a time after the collision are needed to infer position and momentum for *a* and *b* at a time prior to the collision. In addition, it must also be assumed that no collisions with other objects occur. This result contradicts the overdetermination thesis.

Some readers will have noticed what the more fundamental problem is which underlies the preceding consideration concerning conjunctive causal scenarios: classical collision mechanics is time-symmetric. Time symmetry of a physical theory implies that there is no principal difference between inferring future states of a system from the present state and the opposite operation of inferring past states from the present. If there is no overdetermination in one direction, there is no such overdetermination in the other. If there happens to be overdetermination in one direction, there must be overdetermination in the other direction as well.

Frisch (2005, Ch. 7) and Elga (2001) have pointed out that Lewis's overdetermination thesis is in conflict with the time symmetry of certain microphysical theories, specifically the time symmetry of classical mechanics and the fact that the equations of electrodynamics are time-symmetric. Frisch (2014, p. 204) concludes that Lewis's overdetermination thesis is 'provably wrong'. In response to Elga (2001), Loewer (2007) develops an entropy account of the asymmetry of counterfactuals, which may be extended to an entropy account of causation. These accounts are not applicable to microphysics, at least not directly. Frisch (2014, Ch. 8) described counterexamples to the entropy account of causation at the level of macrophysics.

Our discussion of Lewis's fork theory complements critical work on the overdetermination thesis by Price (1996, Ch. 6) and Price (1992). His key claim is that there are not enough asymmetric forks to ground the asymmetry of all causal relations. Put differently, there is a reason why Reichenbach's original fork theory is impure in the sense that it does not attempt to ground all causal relations by embedding the relation in an asymmetric fork. More specifically, Price argued convincingly that asymmetric forks disappear at the micro level of physical systems. This criticism parallels the arguments by Elga (2001) and Frisch (2005, Ch. 7) against Lewis's overdetermination thesis.

9 The Independence Theory

Hausman's (1998) independence theory attempts to characterize the direction of causation without the Humean convention. The core of this theory is simple: any effect has at least two causes, and these causes are causally independent in the sense of not having a common cause. We leave out further details since they are not necessary for our critical note to follow. Recall that our theory of causation assumes the independence principle for common cause scenarios, as explained in Section 6 of Chapter 8. But we do not think this principle could be used to characterize the direction of causation.

Let us consider a physical system which is described by time-symmetric laws. For simplicity, let's look at a system of classical collision mechanics, and study once more a simple collision. The following figure visualizes the collision of an object a, which is moving, with an object b, which is at rest:

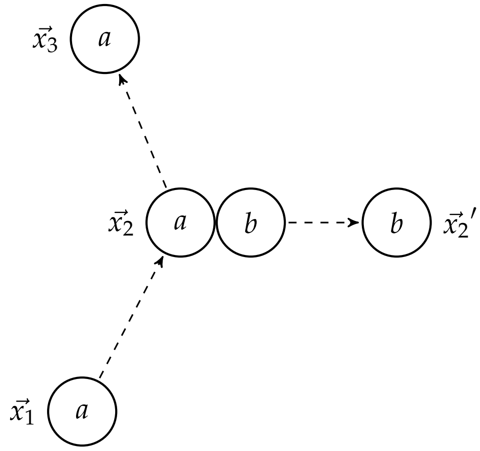


Figure 69: Collision between *a* and *b* at $\vec{x}_2$

Let t_1 be the time point prior to the collision at which a is at position $\vec{x}_1$ and b at position $\vec{x}_2$. t_2 is the point of the collision, and t_3 a time point after the collision. a is at position $\vec{x}_3$ and b at position $\vec{x}_2{}'$ then. Clearly, position and momentum of a, and position and momentum of b at t_1 are causes of the collision. Furthermore, the positions and momenta of a and b at t_3 are caused by the collision and the positions and momenta prior to the collision at t_1. By the independence principle, all causes of the collision are independent of one another.

In line with common sense, we take it that causation in classical collision mechanics is forward-directed in time. The arrows in the figure indicate the direction of causation. It holds that $t_1 < t_2 < t_3$. The simple collision may be part of a larger physical system with further objects and further collisions prior to and after that collision. The trajectories and collisions of the system may be represented by a possible world w. For simplicity, we consider the trajectories only for a finite period T of time.

Now, let us consider a physical system w' which is obtained by reversing the time order. Mathematically, this can be done by defining a new time function: $t' = T - t$. If object a is at time t_1 in world w, it is at this position at time $t'_1 = T - t_1$. If object a has the momentum $\vec{p}$ at time t_1, then it's momentum at time $t'_1 = T - t_1$ has the same value but the opposite direction in space.

By time symmetry of the laws of classical collision mechanics, we know that these laws hold in w'. The systems in both w and w' are deterministic: given the positions and momenta of the objects at a certain time point t, the positions and momenta of all objects are determined by the laws for any other time point.

Note that the possible world w' is not only a hypothetical consideration. At least for some physical systems, we can easily set up a system which has the properties of w'. For example, in w', object a moves from position $\vec{x_3}$ to position $\vec{x_2}$. Likewise, object b moves from position $\vec{x_2}'$ to position $\vec{x_2}$ so that the two objects collide in w' as well. They just come from another direction. Hence, we should say that the collision between a and b in w' is caused by the positions and momenta of a and b at time $t_3' = T - t_3$.[11]

By the independence principle, all causes of the collision between a and b in world w' are independent of one another. Such causes do not have a common cause. Notice, furthermore, that the causes of the collision between a and b in w' are just the effects of the collision between a and b in w—with the qualification that the momenta of objects in w' point to the opposite direction in space. Because of this symmetry, it follows from the independence of the causes of the collision between a and b in w' that the effects of the collision between a and b in w satisfy the independence principle as well! By the independence principle, such effects should count as causes of the collision between a and b in w, even though they are in the future of that collision. For this conclusion it is important that the independence principle serves as a sufficient characterization of the direction of causation in the independence theory.

Hausman has basically three options here. First, point out that the independence principle is an important aspect of causation, but insufficient to account for the asymmetry of causation. Second, conclude that, in physical systems which are governed by time-symmetric laws, we have backward causation inasmuch as we have forward causation. This option implies a gross departure from our causal judgements. Any difference between causes and effects seems to disappear for time-symmetric systems then. Third, conclude that we cannot make sense of causation in closed systems which are governed by time-symmetric laws.

[11] Things are less straightforward for classical electrodynamics. The equations of electrodynamics are time-symmetric too. But there are actual systems w such that the system w'—obtained from w by a reversing the direction of time—is not something we can observe in nature. See Frisch (2014, Chs. 5 and 7) for a detailed discussion.

Of course, Hausman (1998, Sect. 7.4n) is aware of the problem, and has a response to it. The response is complex, and we should refrain from giving an oversimplified summary. We understand that Hausman is willing to accept that there is no causation in closed systems which are governed by time-symmetric laws. Relatedly, interventions are considered as a means to recognize an asymmetry between the two worlds w and w'. Interventions are possible only if the system in question is not closed.

The interventionist response to the problem in question requires a reply to Reichenbach's argument that the notion of an intervention already presupposes the Humean convention. We have reviewed and embraced this argument in the above discussion of Woodward's interventionist account. Hausman is well aware of the argument. The discussion of it remains inconclusive, and so Hausman seems to acknowledge that Reichenbach's point is difficult to refute. More recently, Price (forthcoming) has explored the connection between an interventionist approach to causation and the forward-directedness of causation. Unlike Woodward (2003), Price (1992, 1996) has not tried to eliminate the human agent from the interventionist approach to causation.

To our mind, it remains an important desideratum to make sense of causation even for closed systems. The reason for this is as follows. In physics classes and more advanced research in physics, people often study closed systems using idealizations for simplicity. Otherwise, the mathematical description is simply not manageable, even with our most advanced computational devices. The ideal gas model is a case in point. We want to say, for example, that the pressure of an ideal gas is caused by the collision of the gas molecules with the walls of the container. Even in theoretical meteorology, people are studying closed systems, while knowing that the atmosphere of the Earth is not closed at all. Our causal judgements in science encompass causal relations in closed idealized systems, some of which are governed by time-symmetric laws.

A final note on the refined INUS account by Baumgartner and Falk (2019) is in order. This account runs into an analogous problem. Clearly, we think that the positions and momenta of the two objects at t_1 in world w are causes of their collision. This causal judgement should, in some way or other, be captured by an INUS-style biconditional. So these positions and momenta should be conjuncts of a cluster of causal factors in the sense of the INUS account. The cluster may also contain conjuncts which state that no other collisions are happening to the objects between t_1 and t_2. Of course, t_1 and t_2 are mere placeholders for two time points of a collision. Likewise for the names a and b.

Now, recall that an INUS-style biconditional, basically, needs to satisfy just two constraints: it needs to be extensionally true and satisfy a certain condition of non-redundancy. Both the original INUS account by Mackie (1965) and its refinement by Baumgartner and Falk (2019) attempt to analyse causation without references to temporal relations. Since the collision between a and b is determined also by the positions and momenta of a and b at time t_3—and the absence of other collisions in the time period from t_2 to t_3—these positions and momenta should also be part of some cluster of causes in the sense of the INUS account. The corresponding cluster is not redundant. If it was, an analogous problem would arise for the cluster of the forward-directed causes of the collision between a and b. Hence, we have not only forward causation, but also backward causation in classical mechanics.

The proponent of the INUS account has, basically, the same options the proponent of the independence theory has for a response to this problem. First, admit that even a refined INUS account does not suffice to characterize the direction of causation. Second, conclude that there is as much backward causation as there is forward causation in physical systems which are governed by time-symmetric laws. Third, conclude that there is no causation in physical systems which are described by time-symmetric theories. The latter two options do not seem to be appealing. Going for the third option would amount to a partial vindication of Russell's famous (1913) criticism of the notion of cause.

Finally, Noordhof (2020, Ch. 12) discusses at greater length the problem of how to account for the direction of causation in systems which are governed by time-symmetric laws. To address this problem, he takes recourse to a notion of *primitive non-symmetric chance-raising*. From our Humean and post-logical empiricist perspective, such a notion helps very little to resolve the mystery of causation. We found it very difficult to understand how we could determine the direction of non-symmetric chance raising other than by falling back to the Humean convention.

10 The Disjunctive Fork Theory

Lewis's theory of causation may be described as a pure fork theory for the following reason. The direction of all causal relations is thought to be determined by the structure of forks without consideration of temporal relations. The theory was aimed to solve two interrelated problems. First, it promised to be an alternative to the Humean convention which is better motivated and more objective than this convention. Second, it was intended to help with an account of backward causation. However, we have seen that the counterfactual approach to the direction of causation runs into apparently insurmountable problems. For the time being, this account is not a viable option. Let us therefore take a look at impure fork theories in the tradition of Reichenbach (1956).

Dowe (1996) developed an impure fork theory of the direction of causation, which is based on the following observations and assumptions. We certainly do have forks which are open to the future. However, some causal relations may not be a member of an open fork. So let us better not assume that all causal relations are directed through membership in an open fork. We may or may not have forks which are open to the past. If so, we should understand the corresponding causal relations as instances of backward causation. In any case, the overwhelming majority of open forks are open to the future. These ideas are captured by a disjunctive fork theory of the direction of causation.

Explanation 2. Disjunctive fork theory
Suppose we have a net of events such that edges stand for causal relations. None of the edges have a direction. Suppose (C, E) is an edge in the net. Then C is a cause of E iff the edge (C, E) is member of an open fork such that C forms the head and E the tine, or the directed edge (C, E) is aligned with the majority of open forks.

Three points are worth noting. First, Dowe adopts Reichenbach's notion of causal net whose edges are not directed. Dowe's explanation of this notion makes use of temporal relations, without however any presumption about the direction of physical processes. Temporal relations are needed in order to determine whether or not a given edge is aligned with the majority of open forks. Second, an edge (C, E) may be part of two forks such that one fork is open to the future and the other open to the past. If so, we have to conclude that C causes E, and vice versa. Third, the disjunctive approach to the direction of causation does not depend on the transference theory of causation, developed in Dowe (2000). Further details are worked out in Dowe (2000, Ch. 8) and Dowe (1996).

A major objective of Dowe's impure fork theory is to give empirical content to the claim that there is backward causation. On this theory, we have backward causation in our world iff there are conjunctive forks open to the past. The right-hand side of this biconditional has empirical content, provided the notion of open conjunctive fork has such content. Dowe (2000, Ch. 8) works with Reichenbach's statistical notion of conjunctive fork. He does not claim that there is backward causation in our world. Whether there is backward causation rather remains an open question. The disjunctive approach may help answer this question.

Dowe's impure fork theory is motivated by the entanglement of measurements in quantum mechanics, explained in Section 5. Recall that some physicists and philosophers of physics have envisioned backward causation along the lines of the following figure:[12]

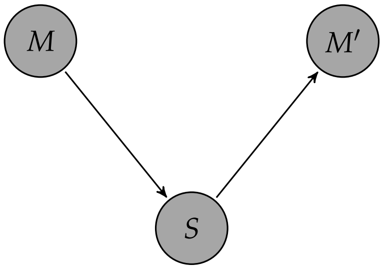


Figure 84: Backward causation in a scenario of quantum entanglement

The idea is that the outcome of measurement M causes certain hidden variables at the source S to assume certain values, which in turn are causal factors for the outcome of measurement M'. Hence, the correlation between the results of M and M'. S stands for the source from which the two particles are emitted. M and M' stand for the measurement of a quantum-mechanical property. The phenomenon of quantum correlations has been

[12] This figure is just a variant of Figure 73. For consistency with our discussion of Reichenbach's fork theory, we have now adopted Reichenbach's convention that the direction of time goes from the bottom to the top of the page.

experimentally confirmed for the spin of two particles which are emitted from the same source.

Since, however, presumed causal factors at the source S are not accessible independently of at least one measurement, the backward interpretation of entanglement of quantum states remains a mere interpretation so far. At this point, Dowe's disjunctive approach to the direction of causation comes into play. It says that the backward interpretation is justified iff we can find another event X in the past of measurement M such that XMS forms a fork open to the past. Such an event must be identified at the type level. That is, it occurs whenever the measurement M is carried out. But X doesn't occur otherwise. The corresponding causal graph is depicted by Figure 85:

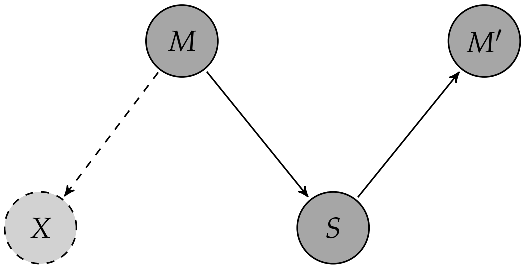


Figure 85: Backward causation: the fork XMS is open to the past.

The disjunctive approach remains non-committal as to whether there is backward causation in our world. It merely specifies empirical conditions under which a backward interpretation of causal processes is justified. From the perspective of our investigation, it is striking that the disjunctive approach can be relatively easily built into our reductive theory. For this to be seen, note that alignment with the majority of open forks is extensionally equivalent to the Humean convention in our world. This criterion of alignment is even intensionally equivalent to the Humean convention if we adopt Reichenbach's definition of what may be called *fork time*: '[I]n a conjunctive fork ACB which is open to one side, C is earlier than A or B' (Reichenbach 1956, p. 162). Of course, this definition needs modification if we want to allow for some forks which are open to the past. The modification is straightforward, though.

To make use of the disjunctive approach, it suffices to note that the Humean convention is extensionally equivalent with the requirement that a causal relation be aligned with the majority of open forks. This equivalence holds at least in our world. With the equivalence in mind, we can generalize our analysis of causation from the previous chapter to make room for backward causation.

Definition 32. Cause
Let C and E be events. C causes E—relative to an epistemic state S—iff

(1) $C, E \in K(S)$

(2) $C \gg_{FN} E \in K(S)$

(3) The directed edge (C, E) is aligned with the majority of open forks, or a member of such a fork with C as head, or C is explanatorily prior to E.

For the sake of conceptual unification, we may furthermore replace the Humean convention with the condition of alignment in our reductive analysis in Chapter 8, which is needed in the definition of explanatory priority. Finally, we need to generalize the notion of a forward-directed deduction. The basic idea of the original definition is that no inferential step goes against the presumed direction of causation. We can retain this idea as follows. Suppose a literal L_E has been directly inferred from a set P of premises in this deduction. We say that this inferential step is in line with the direction of causation iff for all events and absences L_C asserted in P, the directed edge (L_C, L_E) is aligned with the majority of open forks or it is itself a member of an open fork such that L_C is the head event. With this requirement we can generalize our notion of a forward-directed deduction in such a way that temporally backward-directed inferences are admitted.

There remain two problems to be acknowledged before concluding this section. First, the undirected edge (S, M) in Figure 85 is a member of two open forks: one is open to the past and another is open to the future. Taking the disjunctive approach to the direction of causation seriously, we have to conclude that the values of hidden variables at S cause the outcome of measurement M, and vice versa. So we have a situation where C causes E, while E also causes C. Dowe (2000, p. 205) is prepared to accept this conclusion. This may or may not be a severe problem. Dowe points out that we find symmetric causal relations in concrete models of backward causation in physics.

Second, we must wonder whether the second effect of M—simply called X—in the scenario of quantum entanglement is epistemically accessible independently of M and M'. If so, the problem arises that the bilking argument seems to apply to this effect. There is consensus among Hausman (1998), Price (1996), and Dowe (1996, 2000) that the values of the hidden variables at the source are to be epistemically inaccessible—in the absence of the two measurements M and M'—in order to block the bilking argument. By analogy, this consideration also applies to the second effect to be discovered. But if effect X is just as inaccessible as the values of the hidden variables at the source, then it remains challenging to empirically discover instances of backward causation. In his review of Dowe (2000), Hausman (2002) points out this problem.

Now, Dowe (2000, p. 207n) is aware of the problem and has a response to it. He argues that, in order to recognize event X as an effect of the measurement M, we also need to have information about the event of emission at the source S, including the values of hidden variables. And these values are not accessible to us prior to the measurements M and M' and their outcomes. Hence, bilking is blocked. We are hesitant to follow this response. Notably, there is an accessible part of the source S of emission. Let us call S_a the accessible part of S, and S_c the complete event, including the values of hidden variables. Then both XMS_a and XMS_c form an open conjunctive fork. The correlations between X and measurement M is strict. That is, whenever X occurs, measurement M is carried out. Hence, on the basis of observing X and S_a—prior to carrying out M—we can infer that X is caused by the future event M. So, bilking should be possible after all.[13]

In this section, we have outlined a generalization of our reductive analysis which allows for instances of backward causation. The main objective was to show that there is such a generalization which is in line with at least one account of backward causation to be found in the literature. We are hesitant to adopt the present generalization of our analysis for three reasons. First, for the time being, we are lacking empirical evidence for backward causation. Second, there remains at least one open problem of the disjunctive fork theory. Third, so far, we would have to work with a probabilistic notion of a conjunctive fork. It remains an open problem to translate this notion into our inferential framework.

[13] Even if the correlation between X and M was not strict, we can make a probabilistic prediction of M on the basis of X so that bilking seems to be feasible.

11 Conclusion

In this chapter, we have pursued two interrelated objectives. First, to make ourselves familiar with selected attempts at an explanation of backward causation. Second, to review a selection of prominent alternatives to the Humean convention. At the end of this review, we can conclude that no such alternative has succeeded yet. The difficulties and problems arising from alternatives to the Humean convention rather suggest a reconsideration of it.

To support these conclusions, let us try to give a more systematic account of the different types of alternatives and modifications of the Humean convention. We suggest the following classification:

(1) Clear-cut alternatives: the Humean convention is abandoned completely.

(2) Disjunctive approaches: the convention continues to serve as one out of two means to distinguish between causes and effects.

(3) Refinements and liberalizations: the convention is refined in such a manner that the resulting analysis solves the problems of simultaneous and spurious causation.

Lewis's (1973b, 1979) counterfactual theory of causation is perhaps the most prominent clear-cut alternative to the Humean convention. We have seen that it runs into at least two severe problems. It is, in principle, unable to account for the direction of causation in physical systems which are governed by time-symmetric laws. Attempts to improve Lewis's approach by reference to the growth of entropy over time may work at the macro level, but fail to work at the micro level. Moreover, Lewis's counterfactual theory implies that there is unlimited exponential growth of concurrent events. This is highly implausible, and at odds with our observations and accounts of at least a number of physical systems.

Woodward's interventionist account of causation makes use of causal graphs, but avoids references to the Humean convention. We have shown that the direction of causation remains indeterminate for causal relations concerning past events on Woodward's interventionist account—unless the Humean convention is adopted. This result is based on Reichenbach (1956).

Moreover, we have shown that the bilking argument against backward causation goes through on Woodward's interventionist account of causation. This type of causation is not a concern for Woodward (2003), though.

Two more clear-cut alternatives to the Humean convention have been considered: the independence theory by Hausman (1998) and the refined INUS account by Baumgartner and Falk (2019). Both are unable to account for the direction of causation in physical systems which are governed by time-symmetric laws. Hausman (1998) is well aware of the problem, and is inclined to admit that we cannot make sense of causation in closed physical systems which are governed by time-symmetric laws. We have argued that this conclusion is not appealing since people in physics and meteorology often times study such systems. At least sometimes we want to make causal claims for time-symmetric systems.

Another response to the problem of time-symmetric systems is to accept that there is as much forward causation as there is backward causation in such systems. If C is a cause of E, E is a cause of C too. This implies that there is no difference between causes and effects in time-symmetric systems. Again, this response is hardly appealing. It would rather support Russell's (1913) scepticism about causation.

Let us now move on to the disjunctive approaches. Dowe (2000) draws on Reichenbach's original fork theory, and proposes, roughly, the following account of the direction of causation: C is a cause of E iff the ordered pair (C, E) is a member of an open fork, or aligned with the majority of open forks. It is easy to show that the latter criterion is, at least extensionally, equivalent with the Humean convention. Dowe set forth the disjunctive approach as an attempt to make sense of backward causation. Without actually adopting this approach, we have outlined how it could be embedded into our theory. This way, we could account for the conceptual possibility of backward causation, while the Humean convention continues to do almost all of the work when it comes to distinguishing between causes and effects. However, there remains to meet the challenge of the bilking argument.

Reichenbach's (1956) original fork theory should not be omitted in this conclusion. As the title of his seminal work implies, Reichenbach's primary concern is with the direction of time. The direction of causation falls into place once the direction of time is explained. Reichenbach thinks and argues that the former direction is strictly aligned with the latter. He basically thinks that the Humean convention is part of our concept of causation. As for the direction of time, his account is disjunctive.

To be precise, Reichenbach proposes at least two accounts of the direction of time. One is based on the growth of entropy. Another is based on the notion of an open fork. In essence, event A precedes event B iff the ordered pair (A, B) is a member of an open fork such that B is the tine event, or this ordered pair points in the same direction as the other open forks. It is assumed that all open forks point in the same direction in time. Since the direction of causation is strictly aligned with the direction of time, Reichenbach's approach to the direction of causation may be said to be disjunctive in nature too.

Finally, it is striking that Price's (1996) account of backward causation is embedded in a disjunctive approach to the direction of causation as well. The Humean convention remains one of two means to distinguish between causes and effects. For this to be seen, recall that Menzies and Price (1993), and Price (1996) set forth an interventionist account of causation. Such an account rests, in part, on counterfactuals. Menzies and Price are well aware of this and discuss counterfactuals at greater length. Finally, recall that Price (1996, Ch. 7) maintains a weakened version of the Humean convention for the semantics of counterfactuals: when evaluating a counterfactual, we hold the past fixed to the extent it is, in principle, epistemically accessible to us. Hence, the Humean convention determines the direction of causation—for a large range of causal relations—as this convention is built into the semantics of counterfactuals.

A deviation from the Humean convention is only admissible if certain meta-theoretical principles of theory choice favour an interpretation of certain correlations in terms of backward causation. In essence, causation is forward-directed in time unless certain meta-theoretical principles suggest otherwise. Loosely speaking, Price (1996) retains a light version of the Humean convention. And he sees a substantial connection between the convention and human agency (Price forthcoming).

A closer look at the literature on backward causation thus yields a surprising result: prominent, viable accounts of causation which are aimed at capturing both forward and backward causation retain the Humean convention in one form or other. To be precise, this convention or an equivalent convention remains one of two means to distinguish between causes and effects. This result is an important motivation for our reconsideration of the Humean convention.

Our own theory of causation, which is broadly Humean, makes use of the convention only in a modified and refined way. For the time being, we favour the reductive analysis completed in Chapter 9: this analysis solves the problem of spurious causation, and accounts for simultaneous causal relations. We have merely outlined how it may be extended to account for the conceptual possibility of backward causation.

Chapter 11

Conclusion and Synthesis

We set out to analyse causation by way of studying inferential pathways from causes to effects. Thereby, we aimed to reconstruct how a candidate cause brought about a given effect. The reconstruction begins with an operation of suspending judgement about the candidate cause and its effect. A minimum requirement for causation is that—after such a suspension of judgement—the effect can be inferred from the candidate cause together with the laws of the respective background theory.

We have expressed the minimum requirement for causation by an epochetic conditional $C \gg E$. This conditional has the intuitive meaning that C is a reason for E, given C and E are believed. For C to be a genuine cause of E, the proposition that E occurs must be a reason for the occurrence of E. The next step is to impose further constraints on the inferential relations between C and E so as to characterize genuine causal relations.

We have developed two analyses of causation in terms of inferential pathways. Each analysis is centred on two inferential constraints. We will now review very briefly the two analyses, and then show how they can be merged into a coherent and comprehensive theory of causation. Finally, we conclude with an outlook at how our epistemic analysis may be extended to an account of causation in the objects.

1 Causes in Causal Models

The first condition of the causal model analysis requires that each inferential step to a literal, made by a structural equation, must depend on the assumption of the candidate cause. We express this condition by the notion of an active path.

Definition. Active Path
Let $\langle M, V \rangle$ be a causal model, which is uninformative on the literals C and E. There is an active path leading from C to E in $\langle M, V \rangle$ iff E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal—by a structural equation—depends on C.

In less formal terms, existence of an active path means that each inferential step to a literal may be interpreted as a section in a causal process which was started by the candidate cause.

The second condition for genuine causation concerns the notion of deviancy. It has two parts. First, genuine causes are at least weakly deviant. Second, if there are events or absences in the context of the candidate cause which are not deviant, then we must not suspend judgement on them when looking for an agnostic model with an active path. We judge causal relations against a background of weakly normal events and absences if there are any in the context of the candidate cause. Our final definition of an epochetic conditional $\gg$ goes as follows:

Definition. $\langle M, V, N \rangle \models C \gg E$
Let $\langle M, V, N \rangle$ be an extended causal model. $C \gg E$ iff there are $V' \subseteq V$ and $M' \subseteq M$ such that

(1) $\langle M', V' \rangle$ is uninformative on C and E.

(2) There is an active path from C to E in $\langle M', V' \rangle[V']$.

(3) All the structural equations of C's descendants are in M'.

(4) C is weakly deviant and any literal $C' \in V \setminus V'$ which is not a descendant of C in M', and different from C, is deviant.

The notions of deviancy and weak deviancy are explained in terms of inferences from a set N of norms and default laws.

The justification for condition (3) is twofold. First, when suspending judgement on the candidate cause and its effect, all inferential relations between the two are to be preserved. This is why we must not suspend judgement on structural equations of the descendants of the candidate cause. Second, sometimes causes of a common effect are entangled by a causal connection which goes through an ancestor of the candidate cause. Then it may be necessary to disentangle these causes in order to reconstruct a certain causal pathway. Condition (3) permits us to suspend judgement on causal relations among the non-descendants of the candidate cause, and to thereby disentangle the causes of a common effect.

In sum, there are two conditions for actual causation. First, cause and effect occur. Second, the effect is inferable from the candidate cause—after an operation of suspending judgement on both—along an inferential pathway such that each inferred literal depends on the candidate cause, where that cause and its context satisfy a condition of deviancy. By the second condition we aim to capture a notion of factual dependence of the effect on its cause. The two conditions can now be stated by the template of our epochetic approach to causation in causal models.

Definition. Cause
Let $\langle M, V, N \rangle$ be an extended causal model such that $V \models M$. C is a cause of E relative to $\langle M, V, N \rangle$ iff

(C1) $\langle M, V \rangle \models C \wedge E$, and

(C2) $\langle M, V, N \rangle \models C \gg E$.

Note that C and E are literals, which may well stand for absences.

We have shown that our causal model analysis delivers the intuitive verdicts for virtually all causal scenarios which have received some attention in the literature. We are not aware of another analysis of actual causation which is equally comprehensive. Specifically, we have shown that our analysis is more comprehensive than the most advanced counterfactual accounts in Andreas and Günther (2025a). The latter fail to agree with at least some of our causal judgements in scenarios which are captured by our analysis.

2 The Reductive Analysis

The reductive analysis poses a greater challenge since we cannot rely on structural equations anymore. Not surprisingly, this analysis is more involved than the causal model analysis. But the core of the reductive analysis has remained relatively simple and, we think, intuitively accessible. Let us begin with the constraints on the inferential pathways from the candidate cause to the effect—after an operation of suspending judgement on both. Only two constraints are needed.

First, each inferential step to a literal must satisfy a weak formulation of the Humean convention: it must not be backward-directed in time. To be more precise, the inferred literal must not stand for an event or absence which precedes an event or absence asserted by a premise of the inferential step in question. Second, each law used on the inferential path—from the candidate cause to its effect—must be non-redundant. We explain the latter constraint on the basis of a minimalist and syntactic notion of law, without taking any distinction between proper laws of nature and accidental generalizations for granted.

Our notion of non-redundant law is inspired by the best system account, but goes beyond extant formulations of this account. We distinguish between two types of redundancy. Deductive redundancy of a law λ in a set Γ means that it can be inferred from $\Gamma \setminus \{\lambda\}$ by classical, deductive reasoning. Analogously, a law λ is said to be abductively redundant in a set Γ iff λ can be inferred from $\Gamma \setminus \{\lambda\}$ by abductive reasoning. To properly explain the latter notion, we have introduced an inference system of abductive reasoning, which is free of causal notions. Both types of redundancy need to be considered in order to properly discriminate between genuine and spurious causal relations. We have spelled out a criterion of non-redundancy which merges the two types into a unified notion. The resulting analysis has been shown to successfully discriminate between spurious and genuine causes for a wide range of causal scenarios.

Finally, the pair of cause and effect must satisfy a liberalized variant of the Humean convention: the cause precedes its effect or they are simultaneous, while the cause is explanatorily prior to the effect. Explanatory priority is in turn explained in terms of Humean causal relations: C is explanatorily prior to E iff we have a causal Humean explanation of C which is independent of E, but not vice versa. This way, we can account for simultaneous causation on the basis of Humean causal relations.

As for backward causation, we have merely outlined how our analysis may be extended to capture it. The proposal is basically an adoption of Dowe's (1996) disjunctive approach to the direction of causation. Notably, one of the two disjuncts—by means of which we can distinguish between causes and effects—is equivalent to the Humean convention, at least extensionally. Since, however, there remain some open problems surrounding the very notion of backward causation, we are hesitant to further extend our analysis by an account of this type of causation.

In sum, a genuine cause must satisfy three conditions on the reductive analysis. First, trivially, cause and effect occur. Second, the effect can be inferred from the cause such that each inferential step to a literal satisfies two constraints: it is not backward-directed in time and does not use a redundant law as premise. Third, the cause precedes the effect or is explanatorily prior to the latter. In more formal terms:

Definition. Cause
Let C and E be events. C causes E—relative to an epistemic state S—iff

(1) $C, E \in K(S)$

(2) $C \gg_{FN} E \in K(S)$

(3) C precedes or is explanatorily prior to E.

For simplicity, we do not repeat the detailed explanations and definitions of redundancy and non-redundancy here. Note that C and E are literals, which may well stand for absences. Recall that the subscript of the epochetic conditional $\gg_{FN}$ stands for the two constraints just explained: the inferences to literals must be forward-directed in time, and all laws are to be non-redundant.

3 Synthesis

It's time to merge the causal model and the reductive analysis into a coherent and comprehensive account of causation. Note that these two analyses have different virtues. They do not apply equally well to all kinds of causal scenarios. The causal model analysis is aimed at capturing as many scenarios of actual causation as possible. We have shown that it succeeds in this endeavour in Part I. But the causal model analysis is not reductive since we lack a reductive explanation of structural equations. So far, such equations represent elementary causal dependences among events and absences.

The reductive analysis, by contrast, fails to recognize certain subtle relations of actual causation. For example, scenarios for which deviancy and normality matter are obviously not captured. Also, the reductive analysis does not seem to be able to discriminate between genuine and preempted causes. It may or may not be possible to extend the reductive analysis by the requirement that there is an active path from the candidate cause to the effect. The price for such an extension would be additional complexity.

Causal relations in science, however, are better captured by the reductive analysis. The reason for this is that the background theory of the causal scenario in question is not constrained by the logical form of deterministic causal models. At least in principle, the reductive analysis applies to scenarios which are described by highly complex mathematical theories. Obviously, a distinctive virtue of the reductive analysis is that it is reductive.

Let us now merge the reductive analysis with the causal model analysis. The simple idea is to use the reductive analysis as foundation of causal models. Suppose $\langle M, V \rangle$ is a deterministic causal model about a causal scenario. Further, suppose S is an epistemic state which contains beliefs about the same scenario. The epistemic state has the format assumed for the reductive analysis: it is a ranked belief base, no sentence has occurrences of modal or causal notions, etc. Finally, we assume that all literals of the language of M have translations to literals in the language of the epistemic state S.

With this in mind, we define that a causal model $\langle M, V \rangle$ is verified by an epistemic state S iff, roughly, all elementary causal relations on the causal model analysis come out as genuine causal relations on the reductive analysis, and vice versa. This relation holds for all edges of the causal model and all complete valuations which are consistent with M. In more formal terms:

Definition 33. Verification of a causal model
Let $\langle M, V \rangle$ be a causal model and S an epistemic state, as just described.

Let L^*_A be the translation of the literal L_A from the language of M into the language of S. Further let V^* be the set of literals in the language of S obtained by the translations of the set V of literals into the language of M. The epistemic state S verifies the causal model $\langle M, V\rangle$ iff two conditions are satisfied:

(1) for each literal L_A, if $\langle M, V\rangle \models L_A$, then $L^*_A \in K(S)$.

(2) For all complete valuations U which are consistent with M and all edges from A to B in the causal graph of M, the following biconditional holds: L_A is a cause of L_B—relative to $\langle M, U\rangle$—on the causal model analysis without deviancy iff L^*_A is a cause of L^*_B—relative to $\bigcup(S * \bigwedge U^*)$—on the reductive analysis.

The first condition says that all literals which are true on the causal model $\langle M, V\rangle$ are beliefs of the epistemic state S. The second condition applies to all edges of the causal graph of M. Suppose L_A and L_B are true on the causal model $\langle M, V\rangle$, while there is an edge from the variable A to B. Then the second condition says that whenever L_A is a cause of L_B on the causal model analysis without deviancy, L^*_A is also a cause of L^*_B on the reductive analysis, and vice versa. This relation must not only hold for the actual valuation V, but for all complete valuations U which are consistent with M. If L_A is in fact a cause of L_B on a valuation U, we call this causal relation *elementary* since it goes along a single edge in the causal graph of M.

The operation $S * \bigwedge U^*$ stands for the revision of the epistemic state S by some conjunction of the set U^* of literals. $\bigcup(S * \bigwedge U^*)$ gives us the union of the ranks which make up the ranked belief base $S * \bigwedge U^*$. This operation removes the priorities between laws and beliefs about facts so that we obtain what is called a *flat belief base*, that is, a belief base without ranks. Removing these ranks is necessary since otherwise we cannot properly capture causal scenarios of entanglement. The reductive analysis is written for a ranked belief base rather than one without ranks in order to simplify its application. But the ranks are not essential. This may be verified using the more detailed account of belief revision in Appendix B.

Since the reductive analysis does not consider deviancy and normality of an event, we have defined the relation of verification using the causal model analysis without deviancy from Chapter 2. This is not a severe restriction for the following reason. Once a causal model $\langle M, V\rangle$ has been verified by an epistemic state, we can extend this model by a set N of norms and default laws. Below we will state our final, reductive analysis, which exploits the verification of causal models by an epistemic state.

Take the famous rock-throwing example about Suzy and Billy for illustration. A simple epistemic state S which verifies this model would be one which basically has material implications instead of the structural equations of the causal model. Also, information about the temporal relations among the events need to be added so that the elementary causal claims of the causal model can be verified. However, the epistemic state which is to verify the causal model may also be more complex and contain further information. For example, we may have precise information about the velocity of Suzy's rock, the material of the bottle, the spatial relations among the children and the bottle, and so on.

The relation of verification between an epistemic state and a causal model is understood in analogy to the model-theoretic relation of verification. Recall that, in model-theoretic semantics, we say that a certain interpretation, or structure, of a formal language verifies a set Γ of sentences iff all members of Γ are true on this interpretation. Notably, one and the same set of sentences may be verified by a variety of different interpretations. One and the same set of sentences has different structures which verify it. Likewise, a causal model may be verified by a variety of different epistemic states.

Notice, furthermore, that a given epistemic state does not determine a causal model uniquely. Take again the famous rock-throwing scenario. A maximally simple causal model of this scenario contains just the following structural equation: the bottle shatters = Suzy throws a rock or Billy does. The valuation V is such that both Suzy and Billy throw a rock, and the bottle shatters. This causal model is verified by an epistemic state which contains all the information given by the informal story of the scenario. Once judgement has been suspended on the candidate cause and its effect, there is an inferential path from Suzy's throw of a rock to the shattering of the bottle such that the inferences to literals are not backward-directed in time. Also, all laws of such a path are non-redundant relative to the laws of the epistemic state. Likewise for Billy's throw. The latter inferential path is more complex since it involves reasoning by cases, but it satisfies all the constraints of the reductive analysis. The details may be verified by the interested reader.

Now, the epistemic state of the informal story does not only verify the maximally simple causal model, but also the standard model of late preemption. We have explained the latter model in Section 6 of Chapter 3. The standard model considers—unlike the maximally simple model—intermediate events between Suzy's and Billy's throw of a rock, and the shattering of the bottle. For example, the epistemic state of the informal story verifies the causal statement that Billy's throw of a rock is a cause of Billy's rock hitting the bottle on the valuation V which says that Billy throws a rock, while Suzy does not. The verification is straightforward.

Let us also briefly exemplify the distinction between elementary and non-elementary causal relations. Suppose, once more, Billy throws a rock, while Suzy does not. Then there is a relation of actual causation between Billy's throw and Billy's rock hitting the bottle. The causal relation is elementary—on the standard late preemption model—since it goes along a single edge in the causal graph of this model. By contrast, the causal relation between Billy's throw and the shattering of the bottle is not elementary on the standard model since there is no direct connection between the variables of the causal relata.

In sum, the relation of verification between causal models and epistemic states is many-to-many. One and the same causal model may be verified by different epistemic states. Reversely, one and the same epistemic state may verify different causal models.

4 The Final Analysis

We have just observed that a given epistemic state does not determine a unique causal model by the relation of verification. We must therefore wonder which causal models—of all the models which are verified by a given epistemic state—we should use as standard for the determination of causal relations. Is it enough to consider just one causal model—which is verified by the respective epistemic state—to test for relations of actual causation in a given scenario?

The answer to this question is in the negative. The problem with this strategy is that very simple causal models of scenarios of preemption are indistinguishable from causal models of disjunctive scenarios. We have just discussed such a model, which is based on a single structural equation: the bottle shatters = Suzy throws a rock or Billy does. On this model, Billy's throw is not preempted. The standard approach to this problem is to consider more fine-grained models which take intermediate variables into account.

In line with the standard approach, our analysis in Part I works well for causal models of preemption with intermediate variables. But our analysis goes beyond the standard approach in that it provides a solution to the preemption problem without intermediate variables. Take the causal model which is only slightly more complex, and contains the following structural equation: the bottle shatters = Suzy throws a rock, or Suzy does not throw a rock while Billy does. Given that both kids throw a rock, there is no active path from Billy's throw to the shattering of the bottle. Hence, our causal model analysis successfully discriminates between genuine and preempted causes using only a slight refinement of the structural equation from the maximally simple causal model.

In other words, our causal model analysis fails to discriminate between preempted and genuine causes only for the maximally simple causal model. Trivially, any refinement of this model—by considering intermediate variables or just a slight refinement of the structural equation—gives us more information about the causal scenario in question.

Notice, finally, that we can observe the following property of monotonicity for our analysis of actual causation in Part I. Whenever C is considered a genuine cause of E—according to our commonsensical causal judgements—our causal model analysis says so on simple and more complex causal models. The genuine cause in a preemption scenario is a case in point. Some non-causes, by contrast, cannot be recognized as such on certain simple causal models. They come out as non-causes only on more complex causal models. The preempted cause in a scenario of preemption is a case in point, as just explained. With these considerations in mind, we state our final analysis of causation in terms of non-causal and non-modal concepts:

Definition 34. Cause
Let S be an epistemic state, as just described. Let N be a set of norms and defaults. Suppose, finally, that C and E are literals. C is a cause of E iff, for all causal models $\langle M, V \rangle$ which are verified by S, it holds that

(1) $\langle M, V \rangle \models C \wedge E$, and

(2) $\langle M, V, N \rangle \models C \gg E$.

At least from a logical point of view, this analysis is fully reductive. The basis of reduction contains no causal or modal concepts.

The final analysis demands to consider all causal models $\langle M, V \rangle$ which are verified by a given epistemic state in order to show that a candidate cause is genuine. In practice, this is neither feasible nor sensible. It rather suffices to go beyond the maximally simple causal models of a given scenario. This heuristic guideline, however, is difficult to be made more precise. Hence, our final analysis demands consideration of all causal models which are verified by a given epistemic state. While this is a strong idealization with regard to our commonsensical causal judgements, it is in line with our overall methodology which admits of idealizations at the theoretical and logical level.

Admittedly, our extended causal model analysis failed to deliver the intuitive verdicts for some scenarios of overdetermination if the causal model omits certain facts from the informal story of the respective scenario. This problem, however, only arises when at least one of the candidate causes accords with a norm or is present by default. For now, it's an open problem for the above idea that a genuine cause may be recognized for all causal models which are verified by a given epistemic state. More work needs to be done on the notions of deviancy and normality to solve this problem.

Consideration of a range of causal models could be avoided in the final analysis by merging the inferential constraints of the causal model analysis more directly with those of the reductive analysis. For example, it may be possible to define the notion of active path for deductions in propositional and first-order logic.

We decided to not merge the two analyses in a more direct manner so as to separate problems of actual causation from specific problems of a reductive analysis. The benefit of this strategy is a modular architecture of our theory of causation. Different types of problems are addressed by different analyses. Then, at the final stage, the two analyses are merged. But the reductive analysis is not built on top of the non-reductive one.

Let us now resume our discussion of preemption scenarios and spell out how the final analysis applies to such scenarios. A maximally simple model of preemption contains just a single structural equation of the following type: the effect occurs = the preempted cause occurs or the genuine cause does. For the rock-throwing scenario, we have: the bottle shatters = Suzy throws a rock or Billy does. Such a model is not false, but omits important information. Most notably, there is no asymmetry between the genuine and the preempted cause anymore. Hence, we cannot tell which of the candidate causes is genuine and which is preempted.

Now, there are at least two ways to refine the maximally simple model. First, we consider intermediate events. This way, we obtain the standard models for early and late preemption, discussed in Chapter 2. Second, we modify the structural equation of the maximally simple model as follows: the effect occurs = the genuine cause occurs, or the preempted cause occurs while the genuine one does not. We have shown that our causal model analysis properly discriminates between genuine and preempted causes on both refinements of the maximally simple causal model.

With these observations, the application of the final analysis to scenarios of preemption almost falls into place. Suppose the epistemic state contains all the information of the informal story concerning a scenario of preemption. Then both the maximally simple and refined causal models with intermediate variables are verified by the epistemic state. The causal model with the modified structural equation is not, at least not for scenarios of early and late preemption. The problem is that the reductive analysis does not solve problems of early and late preemption. Hence, there is an elementary causal statement—saying that the preempted cause is an actual cause of the effect—which comes out true on the reductive analysis, but not so on the causal model analysis.

Our final analysis is nonetheless able to discriminate between preempted and genuine causes. This is so for the following reason. In order to show that the preempted cause does not qualify as genuine on the final analysis, it suffices to have one causal model such that the model is verified by the epistemic state and the preempted cause does not come out as genuine. Causal models of preemption with intermediate variables satisfy these two conditions. Hence, the preempted cause is not genuine on the final analysis, as it should be. Recall that the final analysis demands that a genuine cause qualifies as such on all causal models which are verified by the respective epistemic state.

The genuine cause, by contrast, comes out as such on both the maximally simple and the more refined causal models of the preemption scenario. Hence, the genuine cause qualifies as such on all causal models which are verified by the epistemic state. This implies that it comes out as genuine on the final analysis, as it should be. Note that the causal verdict concerning the genuine cause tacitly assumes the above principle of monotonicity: if a candidate cause qualifies as genuine on a simple causal model, then it does so on all refinements of that model. For simple scenarios, it may be possible to consider all causal models which are verified by a given epistemic state. For more complex ones, it is not. There remains the challenge of finding a proof or counterexample to the monotonicity principle.

The treatment of scenarios of simultaneous preemption—also referred to as *trumping*—is slightly different. In the standard scenario of trumping, we have information that orders from the major trump orders from the sergeant. We may express this information in our epistemic state S by a sentence along the following lines: the soldier follows the command of the sergeant in the absence of a command by the major. Furthermore, we know that both the major and the sergeant give the command 'Advance!' We also know that the major and the sergeant give a command.

Now, when we test for causation between the sergeant's command and the soldier's advancing on the reductive analysis, the result is in the negative. This is so because, after suspension of judgement on the candidate cause and the effect, we do not know which order was given by major. For this reason, we cannot determine if the soldier will follow the order of the sergeant once we suspended judgement on the soldier's advancing. Hence, our epochetic test fails to verify that the sergeant's command is a cause of the soldier's advancing, as it should be.

This result, in turn, implies that the epistemic state S does not verify a maximally simple causal model of the trumping scenario. Such a model just contains the following structural equation: the soldier advances = the major or the sergeant gives this order. By contrast, the epistemic state S does verify more refined causal models, such as the one with the following equation: the soldier advances = the major gives the order to advance, or else the major does not give an order while the sergeant orders the soldier to advance. We have shown in Chapter 2 that our analysis successfully discriminates between the genuine and the preempted cause on the latter causal model. Hence, our final analysis, stated in this section, is able to make this discrimination as well.

It remains to note that intermediate variables could also be used to solve the standard scenario of trumping. For example, we could introduce a variable for the event that the soldier accepts the commend from the major. Likewise for the sergeant's command. These two events are interrelated in the same way Suzy's rock hitting the bottle and Billy's rock hitting the bottle are. The benefit of using intermediate variables is a more uniform treatment of early, late, and simultaneous preemption on the final analysis.

In sum, our final analysis addresses the problem of model relativity by considering a range of causal models. As is well known, a causal model analysis is always relative to a causal model. The causal verdicts may vary with the causal model considered, as just explained. The question thus arises which causal models should be used to determine causal relations. Our answer to this question is that all causal models should be considered which are verified by the respective epistemic state. Only when considerations of normality and deviancy come into play, relativity to a causal model seems hard to eliminate.

Of course, the final analysis remains relative to an epistemic state. This relativity, however, is less severe for two reasons. First, we have further expressive resources when modelling a scenario in terms of classical logic. Second, given the informal story of a causal scenario, it is easy to judge whether an epistemic state gives us a comprehensive account of the story. It is obvious, for example, that maximally simple accounts of a scenario of preemption leave out some information concerning the candidate causes. So, we should use epistemic states—as standard of verification for causal models—which are comprehensive with regard to the causal scenario in question. We are happy to admit that some model relativity and vagueness remain. These may be eliminated in ways outlined in the next section.

Finally, we should not omit one qualification concerning the claim of reduction in this conclusion. In Section 7 of Chapter 7, we have looked at the conceptual order of judgements about temporal relations in fundamental physics. Some of these judgements are based on causal hypotheses rather than the other way around. The latter hypotheses, however, are inferred from causal judgements for which the standard conceptual order of the Humean convention continues to hold: they are based on temporal relations which are, at least apparently, accessible independently of any causal judgements. The details are complex, and so we have not attempted to make this qualification part of our explicit analysis. After all, the final analysis has the logical form of an explicit definition such that all concepts of the definiens are free of causal and modal notions.

5 Causation in the Objects

We set out to resolve the mystery of causation by way of an analysis of the inferential relations between causal and non-causal statements. Our analysis, however, does not go all the way down to the level of objects, independently of our concepts and beliefs. It's time to outline very briefly how the analysis could be connected to the objects of the real world. Two different approaches may be pursued.

First, we take recourse to the hypothetical assumption of an omniscient epistemic state. Arguably, the epistemic state of an omniscient being gives us access to the world of objects, if only hypothetically. Applied to such a state, our epistemic analysis becomes connected with the objects of the real world. It could then be read as a theory of causation in the objects, if only indirectly. We understand that few readers will be impressed by this manoeuvre. The hypothetical assumption of an omniscient being seems to add even more mystery to causation rather than resolving any mysterious aspect of it.

Reference to an omniscient epistemic state could be defended, though, by Putnam's work on metaphysical realism, contrary to his own intentions. Putnam (1980, 1981), famously, argued that standard realist semantics is just as mysterious as the assumption of an omniscient being. The former tacitly assumes the viewpoint of the latter. The notion of realist semantics is in turn explained by the realist principle that truth is entirely non-epistemic. Truth is one thing, our theories and beliefs are another. There is a clear-cut division between mind and world. Of course, Putnam's arguments have remained controversial, and there is no consensus as to whether they succeed. We leave it to the reader whether she considers reference to an omniscient being a viable option.

Second, the connection to the world of objects could be made through some substantial epistemological theory. We may impose the following constraint on the epistemic state S which is used as foundation of a causal model via of our reductive analysis: the agent of the epistemic state does not only believe the propositions of all sentences $\phi \in K(S)$, but has actual knowledge of these propositions. Then the notion of knowledge is explained by whatever epistemological theory the reader prefers. On this strategy, we take it for granted that knowledge implies truth so that the epistemic state is connected to the objects of the real world. Different accounts of knowledge may be plugged into our epistemic theory of causation in this way.

One challenge arising on the second strategy is that the preferred epistemological theory should not use any causal notions as primitives. Otherwise, the resulting account of causation may be charged with remaining mysterious concerning the causal connections between rational belief and the world. A famous instance of this problem is Kant's *Critique of Pure Reason* (1781/1998). At the beginning of the Critique, Kant uses causal notions to explain how things in themselves are related to an epistemic subject. Then we are told that causation is a category which should only be applied to intuitions and perceptions, but not beyond. If applied to things in themselves, we run into various paradoxes and contradictions. Hence, by his own standards, Kant should not have used causal notions to explain the relation between things in themselves and epistemic subjects. Hume saw an analogous problem arising for his own approach to causation:

> The only conclusion we can draw from the existence of one thing to that of another, is by means of the relation of cause and effect, which shews, that there is a connexion betwixt them, and that the existence of one is dependent on that of the other. The idea of this relation is deriv'd from past experience, by which we find, that two beings are constantly conjoin'd together, and are always present at once to the mind. But as no beings are ever present to the mind but perceptions; it follows that we may observe a conjunction or a relation of cause and effect between different perceptions, but can never observe it between perceptions and objects. 'Tis impossible, therefore, that from the existence or any of the qualities of the former, we can ever form any conclusion concerning the existence of the latter, or ever satisfy our reason in this particular. (Hume 1739/2001, Book I, Part 4, Sect. 2, § 47)

Both Kant's and Hume's system invite scepticism concerning knowledge of external objects. Some readers of Kant responded to this challenge by questioning the dictum of a strict separation between mind and world. Epistemic elements are not only needed to analyse our concept of causation, but also our concept of truth. Putnam's (1992) vision of realism with a human face is an attempt in this direction.

A great deal of mystery remains to be explored. In this book, we have merely tried to resolve some of the mystery surrounding our concept of causation building on a basis of non-causal and non-modal concepts. The final analysis has the logical form of an explicit definition such that all concepts of the definiens are free of causal and modal notions. This is our reductive theory of causation.

Appendices

Appendix A

The Logic of Causal Models

We have defined the notion of cause in terms of inferential pathways from causes to their effects. This approach has been pursued for both the causal model and the reductive analysis. For the causal model analysis, inferential pathways are understood in a system of deductive reasoning with structural equations. We have outlined such a system in Chapter 2. In essence, reasoning with structural equations has been defined on the basis of classical propositional logic with the following qualifications.

There are two elimination rules for the equality symbol $=$, which are analogous to Modus Ponens in classical logic. But there is no introduction rule for this symbol. We have two distinct sets of premises: a set M of structural equations and a set V of literals. To exclude causally backward-directed inferences, we have to intervene on the set M of structural equations with the set V of valuations and the candidate cause C. Otherwise, the inference rules of classical logic for negation, conjunction, and disjunction remain in place.

In this appendix, we show that our system of deductive reasoning with structural equations is in fact sound and complete with respect to the semantics of causal models outlined in Chapter 2. We show this result for sets Γ of arbitrary Boolean propositional formulas. This is a generalization of sets V of literals, which served as premises in Part I. Finally, we show how our propositional causal models may be extended to causal models with non-binary variables.

1 Motivation

From a logical point of view, it is striking that deterministic causal models are introduced very much like a formal logical system. But there are some noteworthy differences. While the semantics of interventionist conditionals uses model-theoretic concepts, we do not have a model-theoretic semantics for inferences in a causal model. Nor do we have a system of natural deduction that allows us to capture such inferences. Halpern (2000) and Briggs (2012) devised axiomatizations of interventionist conditionals using causal models. These axiomatizations, however, do not give us a logic of causal reasoning for drawing inferences from a set of structural equations. They define a logic of conditionals, but not a logic of causal reasoning with structural equations as premises.[1] The notion of a structural equation itself has a syntactic flavour, but its explicit definition in Halpern (2000), and Halpern and Pearl (2005) is a semantic one.

Let us begin with reviewing some basic observations on structural equations from Chapter 2. Suppose $A = \phi$ is a structural equation. If ϕ is given or inferred from other premises, we can infer A from it. This inference has a causal meaning: potential causes of an event are on the right-hand side of a structural equation, while effects are on the left-hand side. The notion of a structural equation stands for a nonsymmetric determination of a variable by the values of certain other variables. And this nonsymmetry is supposed to mirror the nonsymmetry of causal relations. If C is a cause of E, then we cannot infer from this that E is also a cause of C. This contrasts with the symmetry of the biconditional $\leftrightarrow$ and the identity predicate $=$ in classical logic. In the absence of causal cycles, a structural equation always represents an asymmetric determination.

In light of a structural equation being nonsymmetric, we can distinguish between two types of inferences with structural equations. First, inferences from causes to effects, and second, inferences from effects to causes. The latter type of inference is commonly called *abductive*. Our logic of causal reasoning is *forward-directed* in the sense that it is only about inferences from causes to effects.

The logic centres on a syntactic notion of a structural equation. We consider

[1] For an investigation of the relation between conditionals based on causal models and conditionals based on possible worlds, see Halpern (2013) and Huber (2013).

the equality symbol of such an equation a distinct logical symbol, and introduce natural deduction rules for it. These rules are supplemented by a model-theoretic semantics. We introduce the notion of a causal model as a set of structural equations, and describe the interpretation of such a model. The present logic of causal reasoning thus allows us to explain the notions of a structural equation and a causal model in a standard logical format.

One word on the arity of variables in a causal model is in order. We begin with propositional causal models, which are restricted to binary variables. This restriction will be lifted in Section 6 using concepts from many-sorted first-order logic.

2 Propositional Causal Models

Let us first develop an account of causal models which will serve as foundation for the present logic of causal reasoning. In spirit, we very much follow Halpern (2000), and Halpern and Pearl (2005). The main difference to these accounts is that we define the notion of a structural equation in a syntactic manner. Two further differences are noteworthy. First, our formalism does not require any distinction between exogenous and endogenous variables. Second, interventions are defined for arbitrary Boolean formulas, not only for conjunctions of atomic formulas.[2]

Causal models represent law-like relations between events. Some events occur and so are actual, other events do not occur and so are non-actual. Any event occurs or does not. We represent events by propositional variables. The truth value of a propositional variable denotes whether or not the corresponding event occurs. A being true means that the corresponding event occurs, while A being false means that the event in question does not occur.

Let P be a set of propositional variables such that every member of P represents a distinct event. $\mathcal{L}_P$ is a propositional language whose logical symbols are the Boolean connectives. It is defined recursively in the standard way: (i) Any $A \in P$ is a formula. (ii) If ϕ is a formula, then so is $\neg\phi$. (iii) If ϕ and ψ are formulas, then so are $\phi \vee \psi$ and $\phi \wedge \psi$. (iv) Nothing else is a formula.

[2] The latter generalization has already been realized in Briggs (2012). We agree with Briggs (2012) on the general strategy for defining such interventions and corresponding conditionals.

We explain the notion of a structural equation in a syntactic fashion. Let A be a propositional variable of $\mathcal{L}_P$. Let ϕ be a propositional formula of $\mathcal{L}_P$ which has no occurrences of the variable A and which is neither a contradiction nor a logical truth. Then

$$A = \phi$$

is a structural equation based on $\mathcal{L}_P$. Nothing else is a structural equation. The intended meaning of such an equation is that the truth value of ϕ determines that of A. This determination has a causal meaning: it goes from causes to an effect. We shall come to see at a later stage how this directedness of a structural equation is expressed in the formalism. Since $\mathcal{L}_P$ is a propositional language, $A = \phi$ is not a formula of $\mathcal{L}_P$. Having defined the notion of structural equation, we can now define the notion of causal model.

Definition 35. Causal Model M
Let M be a set of structural equations, based on the language $\mathcal{L}_P$. For an $\mathcal{L}_P$ sentence ϕ, $Var(\phi)$ is the set of propositional variables which occur in ϕ. M is a causal model iff it satisfies two conditions:

(1) For any $A \in P$, there is at most one $\sigma \in M$ such that σ has the logical form $A = \phi$.

(2) If $A = \phi$ is a member of M, then there is no ϕ' such that (i) ϕ and ϕ' are (classically) logically equivalent, and (ii) $Var(\phi') \subset Var(\phi)$.

In brief, a causal model M is a set of structural equations such that every propositional variable in $\mathcal{L}_P$ has at most one occurrence on the left-hand side of a structural equation in M. Further, no structural equation in M must have vacuous occurrences of a propositional variable. We call the occurrence of a variable in an equation $A = \phi$ *vacuous* iff the value of ϕ is independent of the value of that variable.

A causal model M is uninterpreted in the sense that there is no valuation of the variables associated with it. A causal models $\langle M, V \rangle$, by contrast, contains a valuation V of the variables in terms of literals. For simplicity, we use the term *causal model* for both interpreted and uninterpreted causal models.

3 Natural Deduction

In this section, we set forth a system of natural deduction for reasoning with structural equations. Suppose Γ is a set of Boolean formulas in the language $\mathcal{L}_P$. Further, suppose M is a causal model, as just defined. Which formulas can be derived from Γ and M? We aim to answer this question by devising a system of natural deduction that defines a relation $\Gamma \vdash_M \phi$.

Obviously, our system needs to include the inference rules of the Boolean connectives: negation, disjunction, and conjunction, including rules for the introduction and elimination of a contradiction $\perp$. We assume the reader is familiar with some formulation of these rules.

There remains the logical symbol $=$ of nonsymmetric determination. This is the genuinely novel logical symbol of causal models. Fortunately, reasoning with structural equations is static in the following sense: we normally do not infer a new structural equation from a given set of structural equations. In light of this, there is no need for an introduction rule of the logical symbol $=$. Another reason for not having such a rule will be given at the end of this section.

Which elimination rules best capture our inferences from a set of structural equations? Let us start with a simple proposal:

$$\frac{A = \phi \qquad \phi}{A}.$$

In words, if $A = \phi$ is a member of M and ϕ can be derived, then A. In addition to this rule, we need an inference rule for deriving $\neg A$:

$$\frac{A = \phi \qquad \neg\phi}{\neg A}.$$

However, this pair of inference rules fails to break the symmetry between the left-hand and the right-hand side of a structural equation. We could still draw inferences from effects to causes, and so go against the direction of causation.

For illustration, let us study the model of a concrete causal scenario: two kids are throwing rocks at a bootle, which shatters eventually. Suzy's rock hits the bottle before Billy's does. For this reason, we think that Suzy's throw is a genuine cause, while Billy's throw is preempted. Following

Halpern and Pearl (2005, pp. 861–4), we model this scenario using the following variables:

- ST: Suzy throws her rock.
- BT: Billy throws his rock.
- SH: Suzy's rock hits the bottle.
- BH: Billy's rock hits the bottle.
- BS: the bottle shatters.

The dependences among these variables may be represented by the following graph:

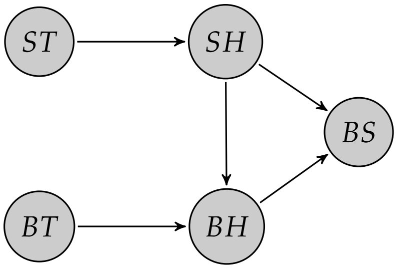


Figure 86: Causal graph of the rock-throwing scenario

The causal model M is given by the following set of structural equations:

$$\boxed{\begin{aligned} SH &= ST \\ BH &= BT \wedge \neg ST \\ BS &= SH \vee BH \end{aligned}}$$

Now, suppose we know that the bottle shattered, but have no direct information about the other events. The premise set Γ is thus given by $\{BS\}$. Then let us start a subproof with the assumption $\neg ST \wedge \neg BT$ (which says that Suzy does not throw her rock and Billy does not throw his either). From this assumption, we can derive $\neg BS$ (which means that the bottle does not shatter). This conclusion contradicts the single member of our premise set $\Gamma = \{BS\}$. By Negation Introduction, we can therefore infer $\neg(\neg ST \wedge \neg BT)$. From this, we can derive $ST \vee BT$, which says that Suzy or Billy throws a rock. Clearly, we have inferred a sentence about potential causes from a sentence about an effect. The above rules fail to express the nonsymmetry of structural equations.

Therefore, we need to impose further constraints on the application of the above rules. Let $Var(\Gamma)$ be the set of propositional variables which occur in at least one formula of Γ, our set of premises. Then we require that a structural equation $A = \phi$ can only be used if A has no occurrences in any formula of Γ:

$$\frac{A = \phi \qquad \phi}{A} \qquad [A \notin Var(\Gamma)] \qquad (= \text{Elim}_1)$$

$$\frac{A = \phi \qquad \neg\phi}{\neg A} \qquad [A \notin Var(\Gamma)]. \qquad (= \text{Elim}_2)$$

This simple constraint on the application of the inference rules for $=$ does the trick. It blocks inferences from effects to causes, but allows causal reasoning from causes to effects. The constraint applies to all inferential steps in the deduction, including the inferences in subproofs.

To resume our running example, suppose once more $\Gamma = \{BS\}$ (which means that the bottle shatters). Then the constraint $[A \notin Var(\Gamma)]$ disallows using the structural equation $BS = SH \vee BH$ in whatever derivation from the premise set $\{BS\}$. For $Var(\Gamma) = \{BS\}$ and BS occurs on the left-hand side of the structural equation in question. Hence, there is no way to infer $ST \vee BT$ from BS. Abductive inferences—which go from effects to causes—are blocked, as desired.

It is worth noting that the constraint $[A \notin Var(\Gamma)]$ translates the operation of an intervention into the language of natural deduction. Recall that, in the standard account, an intervention on a contextualized causal model $\langle M, \vec{u} \rangle$ by $\vec{X} = \vec{x}$ assigns a specific value to the variable X—for all X that are a member of $\vec{X}$—such that any assignment by the function F_X is overruled.[3] Put simply, the structural equation defined by F_X becomes irrelevant once we intervene on X. In a similar vein, the structural equation $A = \phi$ becomes irrelevant if we make an assumption about A in the premise set Γ. This set thus expresses an intervention on certain variables in P, which may well be logically complex in the sense of containing disjunctions and conjunctions.

[3] See Halpern and Pearl (2005, Sec. 2) or Pearl (2009, Ch. 7).

Let LC be the logic whose inference rules are given by the standard natural deduction rules for Boolean connectives, extended by the rules (= Elim$_1$) and (= Elim$_2$). LC simply stands for the *logic of causal models*. The definition of a derivation in LC is now straightforward:

Definition 36. $\Gamma \vdash_M \phi$
Let Γ be a set of $\mathcal{L}_P$ sentences, and let ϕ be such a sentence. M is a causal model, based on the language $\mathcal{L}_P$. Let $\mathcal{L}_{PM}$ be given by the set of $\mathcal{L}_P$ sentences united with M. We say that ϕ is derivable from Γ and M—and write $\Gamma \vdash_M \phi$—iff there is a tree of $\mathcal{L}_{PM}$ sentences which satisfies the following conditions:

(1) The topmost sentences are either in $\Gamma \cup M$ or discharged by an inference in the tree.

(2) The bottommost sentence is ϕ.

(3) Every sentence in the tree, except ϕ, is a premise of an application of an inference rule of LC such that the conclusion of this application stands directly below that sentence.

We write $\Gamma \vdash_M \phi$ instead of $M, \Gamma \vdash \phi$ or $M \cup \Gamma \vdash \phi$ in order to emphasize that M and Γ contain different types of premises. The next step is to introduce the semantics of LC.

Let us finally resume the discussion of an introduction rule for the equality symbol. If such a rule was available, we could derive new structural equations from a given set of such equations. Why is this not desirable? Following Halpern and Pearl (2005, p. 847), we take a structural equation to represent a 'distinct mechanism (or law) in the world'. We further assume that such mechanisms are elementary in the sense that the causal model does not provide us with any information about any submechanism. The structural equations of a causal model are thus comparable to atomic sentences in truth-value semantics: as atomic sentences are used to explain the semantics of logically complex sentences, so are structural equations used to analyse complex causal inferences and judgements. A derived structural equation would not be elementary anymore.

4 Semantics

Recall that a structural equation $A = \phi$ in M simply pairs a propositional variable A with a propositional formula. We can therefore define the semantics of structural equations in terms of the semantics of propositional logic. As is well known, the semantics of a propositional language centres on the notion of an assignment of truth values to the propositional variables. Deviating from the simplified semantics in Chapter 2, we will use truth-value assignments instead of valuations V in terms of literals. The reason for this deviation is to achieve better alignment with the standard format of propositional logic. The equivalence between the two semantics is easy to show.

A value assignment $v : P \mapsto \{T, F\}$ maps the set P of propositional variables to the set of truth values. This in mind, we define what it is for a valuation v to satisfy a structural equation:

$$v \models A = \phi \text{ iff, } v \models_{Cl} A \;\text{ iff }\; v \models_{Cl} \phi. \qquad \text{(Def } v \models A = \phi\text{)}$$

In simpler terms, the valuation v satisfies the structural equation $A = \phi$ iff both sides of the equation have the same truth value on v. Notably, at this stage, the semantics of $=$ does not differ from the semantics of the standard biconditional $\leftrightarrow$ of classical logic. $\models_{Cl}$ stands for the satisfaction relation in classical propositional logic. Deviating from the simpler semantics in Chapter 2, we use truth-value assignment instead of valuations V in terms of literals. The reason for this deviation is to achieve better conformity to the standard format of the semantics of propositional logic.

The satisfaction relation for sets which contain propositional formulas and structural equations can now be defined in the standard way. Let Δ be a set of formulas such that any $\delta \in \Delta$ is either an $\mathcal{L}_P$ formula or a structural equation based on $\mathcal{L}_P$.

$$v \models \Delta \text{ iff, for all } \delta \in \Delta, v \models \delta. \qquad \text{(Def } v \models \Delta\text{)}$$

It seems as if we could define the relation of logical entailment in a straightforward manner as well:

$$\Delta \models \phi \text{ iff, for all } v \text{ s.t. } v \models \Delta, v \models \phi.$$

However, this relation of logical entailment fails to capture the nonsymmetry of the equality symbol in a structural equation. For a valuation v

satisfies a structural equation $A = \phi$ iff v satisfies $A \leftrightarrow \phi$. The two formulas have the same truth conditions.

How can we express the nonsymmetric determination of the equality symbol $=$ within a relation of entailment? Recall that we view the premise set Γ—with regard to the relation $\Gamma \vdash_M \phi$—as expressing an intervention on the propositional variables which have occurrences in Γ. Once we intervene on a variable A, the structural equation determining A—if there is one—becomes irrelevant for the determination of A. Hence, we can say that a possibly complex intervention Γ on the causal model M turns M into a set M_Γ:

$$M_\Gamma = \{\sigma \mid \sigma \in M, \sigma \equiv A = \phi, \text{ and } A \notin Var(\Gamma)\}. \qquad \text{(Def } M_\Gamma\text{)}$$

M_Γ is the subset of M such that $A = \phi$ is in M_Γ iff A does not occur in any formula of Γ. Using this operation on a set M of structural equations, we can define the relation of entailment for our logic LC:

Definition 37. $\Gamma \models_M \phi$
Let Γ be a set of $\mathcal{L}_P$ sentences and let ϕ be such a sentence. M is a set of structural equations, based on the language $\mathcal{L}_P$. We say that ϕ is entailed by Γ and M—and write $\Gamma \models_M \phi$—iff, for all valuations v such that $v \models \Gamma \cup M_\Gamma$, $v \models \phi$.

Notice that the nonsymmetry of $=$ comes into play through an intervention on M, which is expressed by the operation of turning a set M of structural equations into a set M_Γ of such equations. M_Γ is obtained from M by eliminating all structural equations that determine a variable that occurs in a formula of Γ. The union of Γ and M_Γ plays a role that is analogous to the notion of a submodel in Pearl (2009, p. 204).

Let us briefly relate the entailment relation just defined to the notation of the simplified semantics from Chapter 2. Suppose the premise set Γ is a set of literals. Then it holds that $\Gamma \models_M \phi$ iff $\langle M, \Gamma \rangle[\Gamma] \models \phi$. The equivalence is easy to show. The deviation in notation is motivated by two reasons. First, the present notation better conforms to standard logical notations. Second, we want to generalize the semantics of causal models from Chapter 2: Γ may contain propositional formulas other then literals. Note that the simpler semantics in Chapter 2 is contained in the present semantics of causal models.

5 Soundness and Completeness

The relation $\Gamma \models_M \phi$ of entailment is aimed at capturing the relation $\Gamma \vdash_M \phi$ of derivability. Or shall we say that the derivability relation for causal models is aimed at capturing the entailment relation? More important than this question of priority is that semantics and proof theory are in harmony with one another. We show soundness and completeness in this section. Let Γ be a set of $\mathcal{L}_P$ sentences, M be as introduced above, and let ϕ be an $\mathcal{L}_P$ sentence.

Theorem 2. Soundness
If $\Gamma \vdash_M \phi$, then $\Gamma \models_M \phi$.

Proof. Soundness can be proven by induction on the number of inferences in a derivation, as is standard. The inductive step for the inference rules (= Elim$_1$) and (= Elim$_1$) is analogous to the inductive step for ($\rightarrow$ Elim) in proofs of soundness in classical logic. Suppose $\Gamma \vdash_M \phi$.

Induction basis: suppose the number of inferences is zero. Since $\phi \in \mathcal{L}_P$, this implies that $\phi \in \Gamma$. So, the derivation consists of a single formula which is a member of Γ. By the definition of $\models_M$, $\Gamma \models_M \phi$ iff, for all truth value interpretations v, if $v \models \Gamma \cup M_\Gamma$, then $v \models \phi$. Since $\phi \in \Gamma$, it holds that, if $v \models \Gamma \cup M_\Gamma$, then $v \models \phi$. Hence, $\Gamma \models_M \phi$.

Induction step. Suppose we have a derivation of n inference steps. Let Γ' be the union of Γ and the set of assumptions that are not in Γ and so far undischarged. By the induction hypothesis, we know that, for all so far derived sentences ψ, it holds that $\Gamma' \models_M \psi$. We need to show that, for any application of an inference rule of LC, if the next inference consists in inferring δ, then we have $\Gamma' \models_M \delta$. For the inference rules of the Boolean connectives (including $\bot$), this demonstration does not differ from the corresponding inductive step in the soundness proof for a natural deduction system of classical propositional logic. We can therefore focus on the inductive step for the inference rules ($= Elim_1$) and ($= Elim_2$).

Suppose the next inference step (following the n-th step in the derivation) has the form

$$\frac{A = \psi \qquad \psi}{A} \qquad\qquad [A \notin Var(\Gamma)].$$

We need to show that, for all valuations v, if $v \models \Gamma' \cup M_\Gamma$, then $v \models A$. Since there are no inference rules for the derivation of a formula of the type

$A = \psi$, the equation $A = \psi$ is a member of M. Because of the condition $A \notin Var(\Gamma)$, it must even hold that $A = \psi$ is a member in M_Γ. By the induction hypothesis, (i) $\Gamma' \models_M \psi$. Suppose v is a valuation such that $v \models \Gamma' \cup M_\Gamma$. Since the equation $A = \psi$ is a member in M_Γ, this implies that $v \models A = \psi$. Hence, by the semantics of $=$, (ii) A and ψ have the same truth value on the valuation v. Further, we can infer from (i) that (iii) $v \models \psi$. So, ψ is true on v. Obviously, (ii) and (iii) imply that A is true on v. In symbols, $v \models A$. Thus, we have shown that, for all valuations v, if $v \models \Gamma' \cup M_\Gamma$, then $v \models A$. This concludes the inductive step for the inference rule $(= Elim_1)$. The demonstration of the inductive step for $(= Elim_2)$ is analogous.

Note finally that at the end of the derivation, when the last inference step has been completed, $\Gamma' = \Gamma$. All assumptions that are not in Γ must have been discharged. Therefore, by complete induction on the number of inference steps, $\Gamma \models_M \phi$. □

Theorem 3. Completeness
If $\Gamma \models_M \phi$, then $\Gamma \vdash_M \phi$.

The following proof exploits two facts. First, a structural equation $A = \phi$ is satisfied by a truth value interpretation v iff $A \leftrightarrow \phi$ is satisfied by v, where $\leftrightarrow$ has its classical meaning. Second, propositional classical logic whose logical symbols are the Boolean connectives is complete.

Proof. Suppose $\Gamma \models_M \phi$. Hence, by definition of the entailment relation $\models_M$, for all valuations v such that $v \models \Gamma \cup M_\Gamma$, $v \models \phi$. Let M'_Γ be the set that we obtain from M_Γ by replacing every structural equation $A = \psi$ in M_Γ by the classical biconditional $A \leftrightarrow \psi$. Since the truth conditions of $=$ do not differ from the truth conditions of $\leftrightarrow$, $\Gamma \models_M \phi$ implies (i) $\Gamma \cup M'_\Gamma \models_{Cl} \phi$. Further, let M''_Γ be set the set that we obtain from M'_Γ by replacing every biconditional $A \leftrightarrow \psi$ in M'_Γ by $(A \vee \neg\psi) \wedge (\neg A \vee \psi)$. In symbols, $(A \vee \neg\psi) \vee (\neg A \vee \psi) \in M''_\Gamma$ iff $A \leftrightarrow \psi \in M'$. Since $(A \vee \neg\psi) \wedge (\neg A \vee \psi)$ and $A \leftrightarrow \psi$ are satisfied by the same classical valuations v, (i) implies (ii) $\Gamma \cup M''_\Gamma \models_{Cl} \phi$. By completeness of classical propositional logic, this implies that $\Gamma \cup M''_\Gamma \vdash_{Cl} \phi$. Since classical propositional logic with just the logical symbols $\vee, \wedge$, and $\neg$ is complete, $\Gamma \cup M''_\Gamma \vdash_{Cl} \psi$ holds, even if $\vdash_{Cl}$ is defined in terms of the inference rules of the Boolean connectives (including $\bot$). So, (iii) there is a derivation of ϕ from $\Gamma \cup M''_\Gamma$ using only the classical inference rules of the Boolean connectives (including $\bot$).

Now, we can show that $\Gamma \vdash_M (A \vee \neg\psi) \wedge (\neg A \vee \psi)$ for all $A = \psi \in M_\Gamma$. Let us first show that (iv) $\Gamma \vdash_M A \vee \neg\psi$ if $A = \psi \in M_\Gamma$. This can be done by the following derivation:

$$1\;\frac{\overset{\vdots}{\psi \vee \neg\psi} \qquad \dfrac{\dfrac{[\psi]^1 \qquad A = \psi}{A}\,(= \mathrm{Elim}_1)}{A \vee \neg\psi}\,(\vee\text{ Intro}) \qquad \dfrac{[\neg\psi]^1}{A \vee \neg\psi}\,(\vee\text{ Intro})}{A \vee \neg\psi}\,(\vee\text{ Elim})$$

Note that $\psi \vee \neg\psi$ can be derived from the empty premise set using only inference rules of *LC*, which is indicated by the vertical dots on the left-hand side. Analogously, we can show (v) $\vdash_M \neg A \vee \psi$ if $A = \psi \in M$. From (iv) and (v) we can infer $\vdash_M (A \vee \neg\psi) \wedge (\neg A \vee \psi)$ if $A = \psi \in M$. Thus, we have shown that (vi) $\Gamma \vdash_M (A \vee \neg\psi) \wedge (\neg A \vee \psi)$ for all $A = \psi \in M_\Gamma$.

Recall that we have shown that (iii) there is a derivation of ϕ from $\Gamma \cup M''_\Gamma$ using only the inference rules of the Boolean connectives (including $\bot$). By (vi), we know that we can transform this derivation into a derivation of ϕ from Γ and M in the logic *LC*. The transformation goes as follows: instead of taking the sentences $(A \vee \neg\psi) \wedge (\neg A \vee \psi)$ in M''_Γ as premises, we derive these sentences from the structural equations in M using the inference rules of *LC*, as just demonstrated. Hence, $\Gamma \vdash_M \phi$.

□

6 Non-Binary Variables

So far, we have studied propositional causal models. In such a model, all variables are binary. Let us now lift this restriction, and define causal models with non-binary variables. For this to be achieved, some elements of first-order logic are needed. What is a non-binary variable in a first-order language? Which causal scenarios require non-binary variables? Let us study a simple and well-known example. A tower of a certain height casts a shadow. The length of the shadow causally depends on the height of the tower and the angle of the sun rays. These quantities may be represented by first-order functions. More precisely, they can be represented by the values of unary functions for certain arguments:

- a: tower
- b: sun rays
- c: shadow
- $h(a)$: height of the tower
- $n(b)$: angle between sun rays and surface of the earth
- $l(c)$: length of the shadow.

The values of these functions are governed by the following equation (in which *cot* stands for the trigonometric function of cotangent):

$$l(c) = cot(n(b)) \cdot h(a).$$

This equation can be read as a structural equation. For we think that the length of the shadow causally depends on the height of the tower and the angle of the sun rays, but not vice versa. So let us take the equation as a structural equation in the technical sense of causal models. Also, let M be the causal model that contains only this equation.

On this reading, $l(c), n(b)$, and $h(a)$ are the variables of M. It is important to note that the variables of a causal model in a first-order language are not variables in the sense of first-order logic at all. The variables of a deterministic causal model are rather *ground terms* with occurrences of a function symbol. That is, they are terms (in the sense of first-order logic) which do not contain any variables (in the sense of first-order logic) but at least one function symbol. For clarification, we shall also speak of *causal variables* when referring to the variables of a causal model.

It is important to note that not all ground terms in a first-order theory about a causal scenario have a causal role. Take the ground terms for natural numbers. These terms do neither causally determine other variables nor are they causally determined. Likewise, we do not want to understand the constant symbols for tower, shadow, and sun rays as causal variables. Is the tower a cause of its height? This does not seem correct. Nor is it correct to say that all ground terms with an occurrence of a function symbol are causal variables. The value of $2 + 2$, for example, is not causally determined, and so $2 + 2$ should not be considered a causal variable. Hence, choices are to be made as to which ground terms of a first-order theory are considered variables in the sense of the envisioned causal model. Causal modelling is an art (Halpern and Hitchcock 2010).

When working with a concrete causal model, we may want to say that a certain causal variable has a certain value. For propositional causal models, we can simply assert A if we want to say that the variable A has the Boolean value T. For non-binary causal variables, a direct statement about its value has the logical form $f(c) = c'$, where c and c' are individual constants. Note that the equation symbol in such a sentence does not have a causal meaning. If we say that the tower has a certain height, expressed by a rational number and a unit of length, we are thereby not implying that a rational number causally determines the height of the tower. Our logical account of causal models with non-binary variables must therefore distinguish between two equality symbols, one with a causal meaning and another without such a meaning. Let us adopt $:=$ for the equality symbol with a causal meaning. The above structural equation must then be rewritten as follows:

$$l(c) := cot(n(b)) \cdot h(a).$$

If the causal model contains non-causal mathematical equations, these need to be written with the standard equality symbol $=$, not with $:=$.

Each causal variable in a causal model has a well-defined range of values. We can take this into account by working with many-sorted first-order logic. The distinctive feature of this logic is that we have several domains of interpretation instead of a single domain. The different domains correspond to different *sorts*. Any constant symbol must be of a certain sort.

The formation of atomic formulas is constrained by sorts: if R is a predicate of type $\langle \sigma_1, \ldots, \sigma_n \rangle$, then $R(t_1, \ldots, t_n)$ is a formula iff, for all i $(1 \leq i \leq n)$, t_i is a term of sort σ_i. In what follows, let $D(\sigma_i)$ be the domain of interpretation of sort σ_i. The type of a function f is of the form $\langle \sigma_1, \ldots, \sigma_n \rangle \mapsto \sigma_j$. That is, such a function is a mapping of the set $D(\sigma_1) \times \ldots \times D(\sigma_n)$ onto the set $D(\sigma_j)$. Obviously, if f is of the form $\langle \sigma_1, \ldots, \sigma_n \rangle \mapsto \sigma_j$, then $f(t_1, \ldots, t_n)$ is a term iff, for all i $(1 \leq i \leq n)$, t_i is a term of sort σ_i. The term $f(t_1, \ldots, t_n)$ itself is of sort σ_j.

The semantics of many-sorted first-order logic generalizes the semantics of first-order logic in a straightforward manner. An interpretation of a many-sorted language must respect that each constant c is of a certain sort σ_i such that c is interpreted in the domain $D(\sigma_i)$. Likewise for predicates, function symbols, and variables.[4]

Many-sorted first-order logic has been argued to be reducible to standard first-order logic. The precise meaning of this reduction is not obvious, though.[5] In any case, it seems obvious that many-sorted logic allows us to define the range of non-binary causal variables in a relatively straightforward manner. Suppose $f(t_1, \ldots t_n)$ is a ground term and a causal variable of a causal model. Let f be of the sort $\langle \sigma_1, \ldots, \sigma_n \rangle \mapsto \sigma_j$. Then $D(\sigma_j)$ is the range of the causal variable $f(t_1, \ldots t_n)$.

We are now in a position to generalize our account of propositional causal models to causal models with non-binary variables. Let $\mathcal{L}$ be a many-sorted language of first-order logic. Further, let $\mathcal{V}$ be a set of ground terms of $\mathcal{L}$. The members of $\mathcal{V}$ are considered variables of the respective causal model. A structural equation is a sentence of the logical form

$$t := t'$$

where t and t' are ground terms. Moreover, $t \in \mathcal{V}$ and t' must have at least one occurrence of a ground term in $\mathcal{V}$. No other formulas are structural equations of $\mathcal{L}$ with the set $\mathcal{V}$ of causal variables. Note that a structural equation thus defined has no occurrences of quantifiers or first-order variables.

Definition 38. Causal Model (non-binary variables)
Let M be a set of structural equations, based on the language $\mathcal{L}$ with the set $\mathcal{V}$ of causal variables. For an $\mathcal{L}$ term t, $Var(t)$ is the set of causal variables that occur in t. M is a causal model iff it satisfies two conditions:

(1) For any $t \in \mathcal{V}$, there is at most one $\sigma \in M$ such that σ has the logical form $t := t'$.

(2) If $t := t'$ is a member of M, then there is no t'' such that (i) $t' = t''$ on all interpretations of the language $\mathcal{L}$ and (ii) $Var(t'') \subset Var(t')$.

In brief, a causal model based on $\mathcal{L}$ is a set of structural equations such that any term of $\mathcal{V}$ occurs at most once on the left-hand side of an equation in M. Further, no structural equation in M must have vacuous occurrences of a causal variable.

[4] See Enderton (2001, Sect. 4.3) for a textbook account of many-sorted first-order logic.

[5] See Barrett and Halvorson (2017) for a detailed discussion.

Note that a structural equation in a causal model thus defined has no occurrences of quantifiers or first-order variables. Causal models with non-binary variables may or may not be embedded into full first-order reasoning. In this chapter, we merely describe the core of causal reasoning with non-binary variables which is based on a fragment of many-sorted first-order logic. Inference rules for quantifiers do not belong to this fragment.

As for the elimination rule of :=, we adopt:

$$\frac{t := t' \quad t' = t''}{t = t''} \qquad [t \notin Var(\Gamma)] \qquad \text{(Elim :=)}$$

where Γ is the respective set of premises. This inference rule captures the interplay of reasoning about non-causal equality statements and structural equations. There is no need for an introduction rule of := for reasons explained in Section 3. $Var(\Gamma)$ is the set of causal variables of M that occur in at least one premise in Γ.

This is the only genuinely causal inference rule needed in our account of causal models with non-binary variables. All the other inference rules are adopted from classical logic. At the very least, we need the introduction and elimination rules for the non-causal equality symbol =. Once the set of classical inference rules is specified, the definition of $\Gamma \vdash_M \phi$ in Section 3 can be generalized to causal models with non-binary variables in a straightforward manner.

It remains to specify the semantics of causal models with non-binary variables. Recall that we defined the semantics of propositional structural equations $A = \phi$ in terms of classical interpretations of a propositional language. Likewise, we can define the semantics of a structural equation $t := t'$ in terms of classical interpretations. Let $\mathcal{I}$ be a classical, model-theoretic interpretation of $\mathcal{L}$. Then equation $t := t'$ is true on $\mathcal{I}$ iff t and t' designate the same object on the interpretation $\mathcal{I}$.

Note, furthermore, that the set M_Γ (used above in the definition of $\models_M$) remains well defined in the present setting of non-binary variables. M_Γ is simply the set of structural equations of M such that no causal variable of M occurs in any sentence of Γ. Since $\mathcal{V}$ is a set of ground terms, we can even represent an interpretation of M by a set V of literals in a manner analogous to the propositional case. Then the set V_α—which represents the core of V that remains unchanged in an intervention by α—remains well defined too.

Since M_Γ remains well defined, we can easily generalize the definition of $\Gamma \models_M \phi$ in Section 4 to causal models with non-binary variables. We merely need to replace classical Boolean interpretations of $\mathcal{L}_P$ with model-theoretic interpretations of $\mathcal{L}$. Likewise, the proofs of soundness and completeness for causal reasoning with binary variables require only minor modifications to be generalized to the non-binary case. We leave this as an exercise to the reader.

Appendix B

Belief Revision Theory

In this appendix, we define the revision of a ranked belief base in a fully explicit manner. A ranked belief base is one with different levels of priority among its members. Our theory of causation is devised for belief bases which are furnished with such a priority ordering. For this reason, we confine ourselves to studying the revision of ranked belief bases. Then we define an epochetic conditional for ranked belief bases in a more explicit manner. This definition underlies our reductive analysis in Part II.

1 Revision of a Ranked Belief Base

In view of the Levi identity—which says that revisions may be defined in terms of contractions and expansions—we can focus on belief base contractions and thereby define a belief base revision scheme. One way to obtain the contraction of a belief base H by a sentence A goes via the notion of a *remainder set* $H \bot A$. The remainder set $H \bot A$ contains all maximal subsets of H which do not entail A. In formal terms:

Definition 39. $H \bot A$
Let H be a set of sentences and A a sentence. $H' \in H \bot A$ iff

(1) $H' \subseteq H$

(2) $A \notin Cn(H')$

(3) there is no H'' such that $H' \subset H'' \subseteq H$ and $A \notin Cn(H'')$.

Condition (2) is to ensure that no member of the remainder set entails the sentence to be retracted. The rationale for condition (3) is to retain as many of the current beliefs as possible—without retaining the sentence to be retracted—in an operation of contraction. Belief changes are guided by the maxim of minimal mutilation, which goes back to Quine (1961).

Let us now generalize the notion of remainder set to belief bases with a ranking of epistemic priority. Let $\mathbf{H} = \langle H_1, \ldots, H_n \rangle$ be a ranked belief base. That is, $H_1, \ldots, H_n$ are sets of sentences which are explicitly believed, and the indices represent an epistemic ranking of these beliefs. H_1 is the set of the most firmly established beliefs, the beliefs in H_2 have secondary priority, etc. More formally, we can say that the beliefs of a ranked belief base are ordered by a strict weak ordering.[1]

We define the remainder set of a ranked belief base as follows:

Definition 40. $\mathbf{H} \perp A$
Let $\mathbf{H} = \langle H_1, \ldots, H_n \rangle$ be a ranked belief base and A a sentence. $\mathbf{H}' \in \mathbf{H} \perp A$ iff

(1) $\mathbf{H}' = \langle H_1', \ldots, H_n' \rangle$

(2) for all i $(1 \leq i \leq n)$, $H_1' \cup \ldots \cup H_i' \in (H_1 \cup \ldots \cup H_i) \perp A$.

This definition is motivated by the following principle: when retracting a belief A, we should retain the more firmly established beliefs if possible, while the less firmly established ones may be more readily given up. Suppose some beliefs have to be given up because $H_1 \cup \ldots \cup H_n$ entails A. Then, if level i is above level j, we should hold on to beliefs at level i more firmly than to those at level j. For a simple belief base with an upper level of laws and a lower level of presumed facts, this comes down to the requirement that a rational epistemic agent holds on to beliefs about laws more firmly than to those about presumed atomic facts. Suppose we want to contract $\mathbf{H} = \langle H_1, H_2 \rangle$ by A. By assumption, neither H_1 nor H_2 implies A, while $H_1 \cup H_2$ does. Condition (2), then, entails that H_1 is a member of

[1] $<$ is a strict weak ordering iff it is binary, transitive and asymmetric, and the incomparability relation $I(\alpha, \beta) \leftrightarrow \neg(\alpha < \beta) \wedge \neg(\beta < \alpha)$ is transitive. Such an ordering is also called *modular*.

all $\mathbf{H}' \in \mathbf{H} \perp A$. No sentence in H_1 needs to be retracted. But we have to give up at least one sentence in H_2. Consequently, H_2 is not a member of any $\mathbf{H}' \in \mathbf{H} \perp A$.[2]

We are not done yet. The ranked remainder set $\mathbf{H} \perp A$ is not a ranked belief base, but rather a set of such belief bases. To define the contraction of a ranked belief base $\mathbf{H}$ by a sentence A in such a manner that we obtain a ranked belief base, we select a specific member of the ranked remainder set $\mathbf{H} \perp A$. This type of contraction is relative to what's called a *selection function*:

$$\mathbf{H} \div_{\sigma} A = \sigma \mathbf{H} \perp A. \qquad \text{(Def } \div_{\sigma}\text{)}$$

The selection function σ picks out a specific element of the remainder set $\mathbf{H} \perp A$. The contraction operation defined by such a selection function is also referred to as *maxichoice contraction*. Such a contraction allows for a maximally conservative way of retracting a belief: as many of the present beliefs as possible are retained. This behaviour turns out desirable for our analysis of causation, particularly so when we come to analyse a combination of a conjunctive causal scenario with overdetermination. In making the relativization to a selection function notationally explicit, we deviate from standard AGM notations. The selection function, however, is a mere auxiliary concept which will drop out of the final definition of our epochetic conditional.[3]

For the expansion of $\mathbf{H}$, we adopt the following definition:

$$\mathbf{H} + A = \langle \{A\}, H_1, \ldots, H_n \rangle. \qquad \text{(Def +)}$$

Alternatively, we may want to assign a specific level of epistemic priority to the epistemic input A, which would result in an expansion operation indexed by some level i. Such a relativization, however, is not needed for the analysis to follow. Now that contractions and expansions have been

[2] Our notion of a ranked remainder set is inspired by Brewka (1991), but the resulting belief revision scheme is not identical with the one defined in Brewka (1991). While the notion of remainder set is standardly defined for belief sets and flat belief bases, the notion of remainder set of a ranked belief seems to be a genuine contribution to the literature by us.

[3] The expert reader may recall the following problem with maxichoice contractions of belief sets when used to define belief revisions via the Levi identity. For all sentences B, B or $\neg B$ is a member of $K * A = (K \div \neg A) + A$. A belief revision by A, defined by maxichoice contractions, makes us opinionated about everything. This problem, however, does not arise for belief bases.

defined, the definition of belief base revision via the Levi identity falls into place:

$$\mathbf{H} *_\sigma A = (\mathbf{H} \div_\sigma \neg A) + A. \qquad \text{(Def } *_\sigma\text{)}$$

For the belief set $K(\mathbf{H})$ of a ranked belief base $\mathbf{H}$ we define:

$$K(\mathbf{H}) = Cn(\bigcup \mathbf{H}).$$

$K(\mathbf{H})$ is thus given by the classical logical closure of the union of all components of $\mathbf{H}$. Now we can distinguish between belief changes of belief bases and changes of corresponding belief sets. The latter are defined as follows:

$$K(\mathbf{H}) \div_\sigma A = K(\mathbf{H} \div_\sigma A)$$
$$K(\mathbf{H}) + A = K(\mathbf{H} + A).$$

This completes our study of belief changes with an underlying ranked belief base.

2 Epochetic Conditionals

An important lesson of the previous section is that there may well be several ways to contract a belief base by a certain proposition. We have therefore defined a contraction operation which is relative to a selection function, as is standard. Our definition is a generalization of the contraction operation for flat belief bases to belief bases with an epistemic ranking.

The selection function remains implicit in the discussion of all causal scenarios in Part II, simply because there is no need for an explicit discussion. This function is not relevant, for example, to our solution to the problem of spurious causation. We have therefore omitted the selection function in the definition of epochetic conditionals in the main text of Part II.

At the same time, notice that some selection function is present in our non-reductive analysis of causation in Part I. In this analysis, we explicitly take into account that there may be several causal models $\langle M', V' \rangle$ which are uninformative on the candidate cause and its effect. The selection function therefore becomes relevant when we explain how causal models $\langle M, V \rangle$ may be grounded and verified by epistemic states which do not contain any causal notions (see Section 3 in the Conclusion). For this reason, we will now define an epochetic conditional which is relative to a selection function. Then we lift the relativization by existential quantification over all selection functions which satisfy the constraints of the epistemic ranking.

We define a conditional with the following intuitive meaning: $A \gg_\sigma C$ iff, after suspending any beliefs in $K(S)$ as to whether A and C are true or false (using σ), we can infer C from A in the context of the remaining beliefs. In more formal terms:

Definition 41. Belief function $B(A)$
Let A be a sentence and S an epistemic state.

$$B(A) = \begin{cases} A & \text{if } A \in K(S) \\ \neg A & \text{if } \neg A \in K(S) \\ \bot & \text{otherwise.} \end{cases}$$

$$A \gg_\sigma C \in K_>(S) \text{ iff } C \in (K(S) \div_\sigma B(A) \vee B(C)) + A. \qquad (\text{SRT}_\sigma)$$

Equivalently,

$$A \gg_\sigma C \in K_>(S) \text{ iff } (K(S) \div_\sigma B(A) \vee B(C)), A \vdash C$$

where $\vdash$ designates the relation of provability in classical logic. The first step of (SRT$_\sigma$) consists in an *agnostic move* which lets us suspend judgement about the antecedent and the consequent. The contraction by $B(A) \vee B(C)$ gives us an epistemic state in which we do not believe A, B, $\neg A$, or $\neg B$.

Once we have suspended judgement about antecedent and consequent, we check whether or not we can infer the consequent C from the antecedent A in the context of the remaining beliefs of the epistemic state. If so, $A \gg_\sigma C \in K_>(S)$. Otherwise, $A \gg_\sigma C \notin K_>(S)$. $K_>(S)$ is the belief set of the epistemic state S, extended by the Ramsey Test for some conditional. The conditional $\gg_\sigma$ is obviously relative to a selection function σ. To obtain our intended analysis of causation, we eliminate the relativization by means of existential quantification:

$$\begin{aligned} A \gg C \in K_>(S) \text{ iff } & \text{there is } \sigma \text{ such that} \\ & C \in (K(S) \div_\sigma B(A) \vee B(C)) + A. \end{aligned} \qquad (\text{SRT})$$

Suppose now epistemic state S is given by a ranked belief base $\mathbf{H}$. Then we obtain:

$$A \gg C \in K_>(\mathbf{H}) \text{ iff there is } \sigma \text{ s.t. } C \in (K(\mathbf{H}) \div_\sigma B(A) \vee B(C)) + A.$$

This is the definition of our strengthened Ramsey Test upon which our Humean analysis of causation is built. The present definition is to be read as a precise formulation of the corresponding definition in the main text. In the latter definition, we have omitted reference to the selection function σ for simplicity.

Some readers may be interested in the following logical subtleties. First, Gärdenfors (1986) proved a triviality theorem concerning his original formulation of the Ramsey Test in Gärdenfors (1978). Recently, however, there have been various, apparently successful attempts at defending the Ramsey Test in light of this result (see, e.g., Bradley (2007), Leitgeb (2010), and Rott (2011)). Notably, Hansson (1992) has shown how switching from belief sets to belief bases allows us to avoid the triviality result. More specifically, we have shown that our strengthened Ramsey Test avoids Gärdenfors's triviality result in Andreas and Günther (2019).

Second, we use the notation $K(\mathbf{H}) \div A$ as shorthand for *the belief set which results from the contraction of the belief set which is generated by the belief base* $\mathbf{H}$. This notation may be criticized for lacking mathematical rigour. In mathematics, we would assume that $K(\mathbf{H}) \div A$ is well defined for all belief sets K for which there is $\mathbf{H}$ such that $K = K(\mathbf{H})$. The value of a complex function $f(g(a))$ depends only on the object designated by $g(a)$, but not on the object a. This condition is violated for the complex function $K(\mathbf{H}) \div A$ since one and the same belief set may be generated by different belief bases, which in turn yield different contractions by A. The problem may be resolved by defining belief base contractions and revisions as operations on an ordered pair $\langle K(\mathbf{H}), \mathbf{H} \rangle$. For simplicity, we leave the notation as is.[4]

Third, there are epistemic states for which we have both $A \gg B$ and $A \gg \neg B$, as a result of the quantification over selection functions. This property, however, does not result in an inconsistency of our Ramsey Test for epochetic conditionals since we do not say that all members of the logical closure of $K(\mathbf{H}) \cup K_{<}(\mathbf{H})$ are to be accepted.

[4] In his *Textbooks of Belief Dynamics*, Hansson (1999, p. 306) uses a notation similar to ours, while noting that the belief set to be contracted must be understood as generated by a specific belief base. Thanks to David Makinson for communicating the issue to us.

Appendix C

Proofs

1 Chapter 2

Proposition 1. Suppose $\langle M, V \rangle$ is uninformative on the literals L_A and L_B. Further, suppose $\langle M, V \cup \{L_A\} \rangle \vdash L_B$. Then, for any deduction of L_B from $\langle M, V \cup \{L_A\} \rangle$, there is an undirected path $\langle A, D_1, \ldots, D_n, B \rangle$ ($n \geq 0$) of variables such that, if $n > 0$, the deduction contains an intermediate conclusion for each variable $D_1, \ldots, D_n$.

Proof. Suppose $\langle M, V \rangle$, L_A, and L_B are as explained in the proposition. Further, suppose $\mathcal{D}$ is a non-redundant deduction of L_B from $\langle M, V \cup \{L_A\} \rangle$ in our system of natural deduction for causal models. The deduction of L_B, then, consists of several branches. Each branch starts with a premise in $\{L_A\} \cup V$ or an assumption of a subproof, and goes to L_B. Since L_B is not a logical truth, at least one branch must start from a premise in $\{L_A\} \cup V$. Since $\langle M, V \rangle$ is uninformative on L_B, $\mathcal{D}$ must contain at least one branch which starts from L_A. Each branch of $\mathcal{D}$ contains a sequence $\langle L_{D_1}, \ldots, L_{D_k} \rangle$ of inferential steps to a literal such that L_{D_i} is inferred before L_{D_j} iff $i < j$. Suppose the sequence $\langle L_{D_1}, \ldots, L_{D_k} \rangle$ ($k \geq 1$) of inferential steps belongs to a branch in the deduction $\mathcal{D}$ which goes from L_A to L_B. Note that variable D_1 is different from variable A since, by assumption, the sequence does not contain the premise from which the branch starts. It only contains the literals inferred in the branch, omitting the trivial inference to the top element of the branch.

Since the causal model $\langle M, V \cup \{L_A\}\rangle$ may contain the structural equation of A as well as structural equations of variables of the literals in V, we need to consider forward and backward-directed causal inferences. Causally forward-directed inferences are drawn by the two elimination rules for structural equations:

$$\frac{A = \phi \qquad \phi}{A} \qquad\qquad \frac{A = \phi \qquad \neg\phi}{\neg A}.$$

These two inference rules are also used for backward-directed inferences. The pattern of such an inference is as follows: we start a subproof with an assumption ϕ. Then use the classical Boolean inference rules and at least one structural equation to infer a forward-directed conclusion from ϕ. Then we show that ϕ is classically inconsistent with $V \cup \{L_A\}$ or a causally forward-directed conclusion which has been drawn from $V \cup \{L_A\}$ and the structural equations in M. Classical inconsistency means that a contradiction can be derived using just the Boolean inference rules of classical logic. The inference rules for classical implications cannot be used since we have no classical implications in our logic of causal models.

Let us first assume that the deduction $\mathcal{D}$ does not contain proofs by contradiction. We prove by complete induction the following claim for the sequence $\langle L_{D_1}, \ldots, L_{D_k}\rangle$ of inferential steps (which corresponds to a branch from L_A to L_B in the deduction $\mathcal{D}$, as explained above): for each inferential step to L_{D_i}, there is an undirected path between A and D_i in the causal graph of M.

Induction base: we show for D_1 that there is an undirected path between A and D_1. To infer L_{D_1} directly, we need to use the structural equation of D_1, which has the logical form $D_1 = \phi$. To use the structural equation $D_1 = \phi$ for an inference, ϕ or $\neg\phi$ needs to be inferred. Since the sequence $\langle L_{D_1}, \ldots, L_{D_k}\rangle$ corresponds to a branch which starts with the premise L_A and since D_1 is the first literal inferred in the branch, A must occur in ϕ. By our explanation of causal graphs in Section 2, this in turn implies that there is a directed edge from A to D_1 in the causal graph of M.

Induction step: suppose there is an undirected path between A and the variable D_i, where D_i is a variable for which we have an intermediate conclusion in the sequence $\langle L_{D_1}, \ldots, L_{D_k}\rangle$, where $1 \leq i < k$. We need to show that there is an undirected path between A and the variable D_{i+1} in the causal graph of M. The proof of this is analogous to the proof of the induction base. Note that $L_{D_{i+1}}$ is inferred in the branch from L_A to L_B to which the sequence $\langle L_{D_1}, \ldots, L_{D_k} \rangle$ corresponds. To infer $L_{D_{i+1}}$ in a direct deduction, we need to use the structural equation of D_{i+1}, which has the form $D_{i+1} = \psi$. Notice that the inference to $L_{D_{i+1}}$ occurs after the inferential step to L_{D_i} such that there is no inferential step to another literal between these two steps—in the branch from L_A to L_B in the deduction $\mathcal{D}$. This implies that the inference to ψ or $\neg\psi$ relies on L_{D_i} such that D_i occurs in ψ. By our explanation of causal graphs in Section 2, this implies that there is a directed edge from D_i to D_{i+1}. Since we have assumed that there is an undirected path between A and D_i, this implies that there is an undirected path between A to D_{i+1} in the causal graph of M. This concludes the induction step.

We have thus established the claim of Proposition 1 for non-redundant deductions which are free of proofs by contradiction. Let us now lift the assumption that no proofs by contradiction are used.

Induction base: we distinguish two cases: (i) L_{D_1} has been inferred directly, that is, without proof by contradiction. (ii) L_{D_1} has been inferred by a proof by contradiction. Case (i) has just been dealt with.

Suppose we have case (ii). This means that there is a subproof which starts with an assumption ψ and which ends with the conclusion of a sentence equivalent to $\neg L_A$. Since this conclusion directly contradicts L_A, $\neg\psi$ is inferred. $\neg\psi$ has the form of a literal or is used to infer a literal by the classical Boolean inference rules. This literal is the literal L_{D_1} since it is the first literal inferred from A (except for $\bot$). The subproof of the indirect proof has the form of a direct deduction which starts from ψ and ends with a sentence equivalent to $\neg L_A$. Since we have already shown the claim of the proposition for direct deductions, we know that that there is an undirected path between the variable D_1 (which occurs in ψ) and the variable A in the causal graph of M.

Induction step: suppose there is an undirected path between A and the variable D_i, where D_i is a variable for which we have an intermediate conclusion in the sequence $\langle L_{D_1}, \ldots, L_{D_k} \rangle$ $(1 \leq i < k)$. We need to show that there is an undirected path between A and the variable D_{i+1} in the causal graph of M. The proof of this induction step is obtained from the proof of the induction base by replacing L_A with L_{D_i} and replacing L_{D_1} with $L_{D_{i+1}}$.

Since $\langle L_{D_1}, \ldots, L_{D_k} \rangle$ represents the inferential steps of a branch from L_A to

L_B in the deduction $\mathcal{D}$, the variable D_k is the variable B. We have thus established the claim of Proposition 1 for non-redundant deductions. Since any redundant deduction can be transformed into a non-redundant one by eliminating inferential steps, this implies the claim of Proposition 1 in general. □

Proposition 2. Suppose $\langle M, V\rangle$ is uninformative on the literals L_A and L_B. Further, suppose $\langle M, V\rangle[V][L_A] \vdash L_B$. Then, for any deduction of L_B from $\langle M, V\rangle[V][L_A]$, there is a directed path $\langle A, D_1, \ldots, D_n, B\rangle$ $(n \geq 0)$ of variables such that, if $n > 0$, the deduction contains an intermediate conclusion for each variable $D_1, \ldots, D_n$. Such a directed path exists in the causal graph of $M_{V\cup\{L_A\}}$.

Proof. Suppose $\langle M, V\rangle$, L_A, and L_B are as explained in the proposition. Further, suppose $\mathcal{D}$ is a non-redundant deduction of L_B from $\langle M, V\rangle[V][L_A]$ in our system of natural deduction for causal models. By Proposition 1, there is an undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ $(n \geq 0)$ of variables such that, if $n > 0$, the deduction contains an intermediate conclusion for each variable $D_1, \ldots, D_n$. We can distinguish two cases: $n = 0$ and $n > 0$.

Suppose $n = 0$. This implies that (i) deduction $\mathcal{D}$ contains no intermediate conclusions concerning variables which are on an undirected path between A and B. It also implies that $\langle A, B\rangle$ is an undirected path. Hence, there is a directed edge between A and B. Suppose, for contradiction, the edge goes from B to A and there is no edge from A to B. Notice now that $\langle M, V\rangle[V][L_A]$ equals $\langle M_{V\cup\{L_A\}}, V \cup \{L_A\}\rangle$, where $M_{V\cup\{L_A\}}$ is obtained from M by eliminating the structural equations of variables which have occurrences in V or L_A. Hence, (ii) deduction $\mathcal{D}$ is a non-redundant deduction of L_B from $\langle M_{V\cup\{L_A\}}, V \cup \{L_A\}\rangle$. Now, the assumption that the directed edge between A and B goes from B to A implies that (iii) the structural equation of A is the only structural equation in M which has occurrences of both variable A and variable B.

(i), (ii), and (iii) imply that deduction $\mathcal{D}$ uses at most the structural equation of A, but no other structural equations. Since, however, the structural equation of A is not a member of $M_{V\cup\{L_A\}}$, no structural equations are used in the deduction $\mathcal{D}$ of L_B from $\langle M_{V\cup\{L_A\}}, V \cup \{L_A\}\rangle$. Hence, $V \cup \{L_A\} \vdash L_B$. Since $\langle M, V\rangle$ is uninformative on the literals L_A and L_B, and since A and B are two different variables, it holds that $V \cup \{L_A\}$ and L_B have no descriptive terms in common. In other words, there is no variable D such that D

occurs in the premises $V \cup \{L_A\}$ and the conclusion L_B. $V \cup \{L_A\} \vdash L_B$, therefore, implies that L_B is a logical truth. This contradicts the assumption that L_B is a literal. Thus we have shown that the directed edge between A and B must go from A to B, and so we have established the claim of the proposition for $n = 0$ and non-redundant deductions.

Let us now deal with the case where $n > 0$. By Proposition 1, we know that there is an undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ $(n \geq 1)$ of variables such that the deduction $\mathcal{D}$ contains an intermediate conclusion for each variable $D_1, \ldots, D_n$. Suppose, for contradiction, there is no directed path from A to B. This implies that the undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ has one of the following properties: (a) It contains a directed edge from D_1 to A, while there is no directed path from A to D_1 with corresponding intermediate conclusions in the deduction $\mathcal{D}$. (b) It contains a directed edge from D_{i+1} to D_i $(i < n)$, while there is no directed path from D_i to D_{i+1} with corresponding intermediate conclusions in the deduction $\mathcal{D}$. (c) It contains a directed edge from B to D_n, while there is no directed path from D_n to B with corresponding intermediate conclusions in the deduction $\mathcal{D}$. (d) we have a combination of (a), (b), and (c).

Let us begin with case (a). This case is almost perfectly analogous to the case where we have $n = 0$ and an undirected path between A and B. Note that (i) the structural equation of A is the only equation which has occurrences of both D_1 and A. Further, (ii) let us assume that the deduction of D_1 from $\langle M_{V \cup \{L_A\}}, V \cup \{L_A\}\rangle$ contains no intermediate conclusions concerning variables which are on an undirected path between A and D_1. The case where there are such conclusions can be reduced to case (b), which will be dealt with below. (iii) $\langle M_{V \cup \{L_A\}}, V \cup \{L_A\}\rangle$ does not contain the structural equation of A. (i), (ii), and (iii) imply that $V \cup \{L_A\} \vdash L_{D_1}$. Suppose, for contradiction, variable D_1 occurs in a literal in V. Then deduction $\mathcal{D}$ is redundant or V is inconsistent. However, we have assumed that $\mathcal{D}$ is non-redundant. And we have assumed that $\langle M, V\rangle$ is uninformative on L_B, which implies that V is consistent. Hence, D_1 has no occurrences in V. By the same line of reasoning, we can show that D_1 has no occurrences in L_A. Hence, D_1 has no occurrences in $V \cup \{L_A\}$. Together with $V \cup \{L_A\} \vdash L_{D_1}$, this implies that L_{D_1} is a logical truth. This, however, contradicts the assumption that L_{D_1} is a literal.

Ad (b). In this case, the undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ $(n \geq 0)$ contains a subsection $\langle D, E, F\rangle$ such that $D \rightarrow E \leftarrow F$. There is no edge from E to

D or an edge from E to F. This implies that the structural equations of D and F cannot be used for a deduction of L_E from L_D or in a deduction of L_F from L_E. Hence, the structural equation of E must be used in a deduction of L_F from L_D via L_E. Let $E = \phi$ be the structural equation of E. Since the edges go from D to E and from F to E, ϕ has occurrences of both D and F.

Further distinctions need to be made. Let us first assume deduction $\mathcal{D}$ does not use reasoning by cases. As regards the structural equation $E = \phi$, we can distinguish two cases. (i) L_D and, possibly, further intermediate conclusions—which concern the direct ancestors of E but not F—determine the truth value of ϕ. (ii) L_D does not determine the truth value of ϕ, even in the context of other intermediate conclusions—which concern the direct ancestors of E but not F.

Suppose we have (i). Then we can infer L_E, the value of variable E. But (i) also implies that the truth value of ϕ does not depend on the value of the variable F, given the causal model $\langle M, V\rangle[V][L_A]$. Hence, we cannot use the intermediate conclusion L_E to infer anything about the value of variable F in the deduction $\mathcal{D}$. This, however, contradicts the assumption that $\langle D, E, F\rangle$ is a subsection of the undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ such that $\mathcal{D}$ contains intermediate conclusions about $D_1, \ldots, D_n$.

Suppose we have case (ii). This implies that we cannot infer L_E from L_D, even in the context of other intermediate conclusions—which concern the direct ancestors of E but not F. Hence, L_E is not an an intermediate conclusion between L_D and L_F—in the deduction $\mathcal{D}$. This conclusion contradicts the assumption that $\langle D, E, F\rangle$ is a subsection of the undirected path $\langle A, D_1, \ldots, D_n, B\rangle$ such that $\mathcal{D}$ contains intermediate conclusions about $D_1, \ldots, D_n$.

Let us now lift the assumption that deduction $\mathcal{D}$ does not use reasoning by cases. Then, by starting a subproof with an assumption of the form L_E, we may be able to infer a sentence L_F by the equation $E = \phi$, L_D, and, possibly, other intermediate conclusions—which concern the direct ancestors of E but not F. This may be feasible since we may be able to infer $\neg L_E$ from $E = \phi$, $\neg L_F$, L_D, and, possibly, other intermediate conclusions in a subproof. Let L_F be the conclusion of such an indirect proof. However, the problem of such a proof by cases concerning E is that we need to complete another subproof which starts with a sentence equivalent to $\neg L_E$, and show that the two subproofs result in the same final conclusion. Let L_G be the final conclusion of the two subproofs.

We show that there is an inferential path from a literal equivalent to $\neg L_E$ to L_G which corresponds to a directed path from E to G in the causal graph so that the claim of the proposition remains true. Suppose, for contradiction, there is no such path. Note, first, that (i) the inferential path of the second subproof cannot go from the variable E to variable F. If this was the case, we would have to infer in the second subproof a literal equivalent to $\neg L_F$. This, however, would imply that the inferences from L_D to L_F and to a literal equivalent to $\neg L_E$ are redundant since we could simply do reasoning by cases with regard to the value of F right away. Recall that we assumed that $\mathcal{D}$ is not redundant.

(i) and the assumption that there is no directed path from E to G with corresponding intermediate conclusions implies that the inferential path from E to G contains a section which goes along an undirected path $D' \rightarrow E' \leftarrow F'$ in the causal graph of M. With regard to this section we run now into the same problem observed on the assumption that no reasoning by cases is used: let $E' = \phi'$ be the structural equation of E'. If $L_{D'}$ and, possibly, other intermediate conclusions determine the value of ϕ', then the value of E' is independent of the value of F'. In this case, we cannot infer anything about F' from the value of E'. If $L_{D'}$ and other intermediate conclusions do not determine the value of ϕ', we cannot infer $L_{E'}$ from $L_{D'}$, even in the context of other intermediate conclusions. In that case, there is no inferential path from $L_{D'}$ to $L_{F'}$ via $L_{E'}$ either. Thus we have derived a contradiction from the assumption that there is no inferential path from a literal equivalent to $\neg L_E$ to L_G which corresponds to a directed path from E to G in the causal graph. This concludes the treatment of case (b).

Case (c) is analogous to case (b). If we have a combination of cases (a), (b), and (c), then consideration of one of these cases suffices to obtain a contradiction. This concludes the proof for $n > 0$.

We have thus established the claim of the proposition for non-redundant deductions. Since any redundant deduction may be transformed into a non-redundant one by eliminating certain inferential steps and corresponding intermediate conclusions, this result implies the claim of the proposition for redundant deductions. Thus we have shown that the proposition is true in general. □

Proposition 3. Let $\langle M, V \rangle$ be a causal model, which is uninformative on C

and E. There is a deduction of E from $\langle M, V \rangle[V][C]$ such that the inferential network of this deduction has the property that all nodes are on a directed path from C to E iff E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal depends on C.

Proof. Subsequent to stating this proposition in Section 5, we made three observations. (i) The inferential networks give us complete information about direct and non-direct dependences among the inferred literals, including dependences with respect to C. (ii) If there is a directed path from literal C to literal D—in a network of a deduction from $\langle M, \emptyset \rangle[V][C]$—then we know that D inferentially depends on C. Finally, (iii) if literal D inferentially depends on literal C, then there is a directed path from C to D in the network of the corresponding deduction. With these observations at hand, the proof is relatively straightforward.

For the forward direction, suppose there is a deduction of E from $\langle M, V \rangle[V][C]$ such that the inferential network of this deduction has the property that all nodes are on a directed path from C to E. By (i) and (ii), this implies that all inferred literals in this deduction inferentially depend on C. Hence, E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal (by a structural equation) depends on C.

For the other direction, suppose E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal (by a structural equation) depends on C. Suppose, for contradiction, that the inferential network of this deduction is not an active path in the sense of Explanation 1. By (iii), this implies that there is an inferential step to a literal (by a structural equation) which does not depend on C. This contradicts our assumption that E can be inferred from $\langle M, V \rangle[V][C]$ such that any inferential step to a literal (by a structural equation) depends on C. Thus we have obtained a contradiction. This concludes the proof of Proposition 3.

□

Proposition 4. Whenever the inferential network of a deduction of E from $\langle M, V \rangle[V][C]$ has the form of a sequence, then this sequence is an active path.

Proof. This proposition holds because, first, it follows from the definition of an inferential network $\langle \mathcal{L}, \mathcal{D} \rangle$ and the definition of an intervention that the inferential network of any deduction of E from $\langle M, V\rangle[V][C]$ does not contain a directed edge which goes to C. Hence, if the inferential network has the form of a sequence, this sequence must start with C. Second, since E is by definition the final conclusion of the deduction in question and since the deduction is assumed to be non-redundant, the sequence must end with E as the last element. So any inferential network—of a deduction of E from $\langle M, V\rangle[V][C]$—which has the form of a sequence, must have the form $C \rightarrow \ldots \rightarrow E$. And such a sequence is always an active path by Explanation 1.

□

2 Chapter 3

Proposition 5. Let $\langle M, V\rangle$ be a causal model which is uninformative on C and E. Suppose there is a deduction of E from $\langle M, V\rangle[V][C]$ such that this deduction has an active path. Then there is such a deduction which is direct with respect to all causal inferences.

Let us first explain the proof idea so as to give an overview. Call a proof *backward-directed* iff it goes against the direction of causation. Such proofs go along directed paths in the causal graph, but against the direction of such paths. By contrast, a proof is called *forward-directed* iff it goes along one or more directed paths in the causal graph. Backward-directed proofs require subproofs, while forward-directed proofs do not. All direct proofs are forward-directed.

We prove a lemma which basically says that any backward-directed proof and any sequence of such proofs can be transformed into a direct proof which is forward-directed. For example, if we can derive L_B from L_A in a backward-directed proof, then we can transform this proof into a direct proof in which $\neg L_A$ is derived from $\neg L_B$. Most importantly, the direct proof coming out of such a transformation uses only inferential steps which are used in the backward-directed proof from which the transformation started. To be precise, all inferential steps to a literal which are causal and which occur in the direct proof have already been used in the backward-directed proof. Recall that an inferential step is causal iff it uses a structural equation directly, or indirectly in a subproof. Figuratively speaking, we can reverse a backward-directed proof of L_B from L_A in such a manner that we obtain a direct proof of $\neg L_A$ from $\neg L_B$. Notably, this reversal reuses certain inferential steps of the backward-directed proof.

By means of this result, we can prove the contrapositive of Proposition 5: if there is no direct proof of the effect which has an active path, then there is no indirect proof of the effect with an active path either. To prove this implication, suppose there is no direct proof of the effect which has an active path. Then suppose, for contradiction, there is an indirect proof of the effect E from the candidate cause C which has an active path. This proof starts with the assumption $\neg E$. From this assumption $\neg C$ is derived. Since the presence of the candidate cause C serves as premise, we have obtained a contradiction. From this contradiction we infer $\neg\neg E$, from which we infer E.

This indirect proof contains a backward-directed derivation: $\neg C$ is derived from $\neg E$. We can turn this derivation into a proper proof of $\neg C$ from $\neg E$ by making $\neg E$ a premise instead of an assumption in a subproof. By the above lemma, the proof of $\neg C$ from $\neg E$ can be turned into a proof of E from C such that all inferential steps which are causal have already been used in the backward-directed proof of $\neg C$ from $\neg E$. Since the latter proof is taken out of the indirect proof of E from C, (i) it holds that all causal inferential steps of the direct proof have been used in the indirect proof of E from C as well. Since we assumed that the indirect proof has an active path, it holds that (ii) all causal inferential steps of the indirect proof depend on the candidate cause. (i) and (ii) imply that the direct proof obtained from the indirect proof has an active path. But this contradicts our initial assumption, according to which there is no direct proof of the effect E which has an active path. This contradiction concludes the proof of the contrapositive of Proposition 5. Proposition 5 itself follows therefrom directly.

For simplicity, we have omitted the possibility of combining direct deductions with indirect ones in order to obtain a deduction of the effect which has an active path. This possibility will be considered in what follows. Let us begin with simple backward-directed deductions which result in a literal and which require only a single inferential path.

Lemma 1. Suppose $\langle M, V\rangle$ is uninformative on the literals L_A and L_B. There is a directed path $\langle B, D_n, \ldots, D_1, A\rangle$ in the causal graph of M_V, but none from A to B. Further, suppose $\langle M, V \cup \{L_A\}\rangle[V] \vdash L_B$ such that there is a non-redundant deduction $\mathcal{D}$ of L_B which contains a sequence of intermediate conclusions $\langle L_{D_1}, \ldots, L_{D_n}\rangle$. Finally, suppose, no other inferential pathways are used in the deduction $\mathcal{D}$. Then $\langle M, V \cup \{\neg L_B\}\rangle[V] \vdash \neg L_A$ such that there is a direct deduction of $\neg L_A$ from $\langle M, V \cup \{\neg L_B\}\rangle[V]$. It is direct with regard to all inferential steps by a structural equation.

Proof. We study the deductive relation $\langle M, V \cup \{L_A\}\rangle[V] \vdash L_B$—rather than the relation $\langle M, V\rangle[L_A][V] \vdash L_B$—in order to prove a general property of backward-directed deductions. Since the causal model $\langle M, V\rangle[L_A][V]$ is obtained by an intervention on $\langle M, V\rangle$ with L_A and V, we cannot draw any backward-directed inferences from L_A in this model.

Note that causally backward-directed deductions go along one or more directed paths, but go against the direction of such paths. They proceed in a stepwise fashion via literals just as forward-directed deductions do. And they rely on one or more indirect proofs. The subproof of such an indirect proof goes in the direction of causation, though. Without loss of generality, we assume that each backward-directed inferential step is made by a separate indirect proof. Let us take a closer look at the details of a backward-directed deduction with the properties assumed in the lemma.

Let $A = \phi$ be the structural equation of A. By assumption, L_{D_1} is inferred from L_A by an indirect proof. Suppose (i) L_A is a positive literal, which means that $L_A \equiv A$. Then L_{D_1} is established in the formal deduction as follows: we assume $\neg\phi$, from which $\neg A$ is derived. This contradicts the premise A. From this contradiction we can infer ϕ via $\neg\neg\phi$. From ϕ we can infer L_{D_1} since this literal is an intermediate conclusion and established by an indirect proof in the deduction of L_B. The last inferential step is made in the context of $M_V \cup V$.

Suppose now (ii) $L_A \equiv \neg A$. To use $\neg A$ in an indirect proof of L_{D_1}, we assume ϕ, from which we infer A, which implies a contradiction with $\neg A$. Thus we can infer $\neg\phi$, from which L_{D_1} is derived since this literal is an intermediate conclusion and established by an indirect proof in the deduction of L_B. Again, the latter derivation may use premises in V and intermediate conclusions from $M_V \cup V$.

Things are analogous for an inferential step from L_{D_i} to $L_{D_{i+1}}$. Let $D_i = \phi_i$ be the structural equation of D_i. If L_{D_i} is a positive literal, we assume $\neg\phi_i$ in an indirect proof. Then infer $\neg D_i$, which contradicts L_{D_i}, and so we can infer ϕ_i and $L_{D_{i+1}}$. If L_{D_i} is a negative literal, we assume ϕ_i in an indirect proof. Then infer D_i, which contradicts L_{D_i}, and so we can infer $\neg\phi_i$. Since by assumption $L_{D_{i+1}}$ is inferred from L_{D_i} by an indirected proof, we must be able to infer $L_{D_{i+1}}$ from $\neg\phi_i$, where premises from $M_V \cup V$ may be used in this inference. There is no other way to infer $L_{D_{i+1}}$ from L_{D_i} by an indirect proof.

Notice that these backward-directed inferential steps are inversions of forward-directed inferential steps. If L_{D_1} is inferred from A, we can first infer $\neg\phi$ from $\neg L_{D_1}$ and $M_V \cup V$ since it holds that $M_V \cup V, \phi \vdash L_{D_1}$. Second, we can infer $\neg A$ from $\neg\phi$ and $A = \phi$. So, it holds that $M_V \cup V, \neg L_{D_1} \vdash \neg A$. Most notably, the inferential step from $\neg\phi$ to $\neg A$ has been used in the subproof of the indirect proof. Likewise, if L_{D_1} is inferred from $\neg A$, then this backward-directed inference uses the forward-directed inference from ϕ to A by the structural equation $A = \phi$ in a subproof.

Things are perfectly analogous for each inferential step from a literal L_{D_i} to a literal $L_{D_{i+1}}$. All these inferential steps are based on a forward-directed inference from ϕ_i to D_i or a forward-directed inference from $\neg\phi_i$ to $\neg D_i$. Since $L_{D_{i+1}}$ is derived from ϕ_i (or $\neg\phi_i$), it holds that $\neg\phi_i$ (or ϕ_i) may be derived from $\neg L_{D_{i+1}}$ and $M_V \cup V$. Hence, there is a deduction of $\neg L_A$ from $\langle M, V \cup \{\neg L_B\}\rangle$ which is forward-directed and contains the following sequence of intermediate conclusions: $\langle \neg L_B, \neg L_{D_n}, \ldots, \neg L_{D_1}, \neg L_A\rangle$.

□

Let us now generalize this lemma, and lift the assumption that the backward-directed deduction consists of just a single inferential path.

Lemma 2. Suppose $\langle M, V\rangle$ is uninformative on the literal L_A and a Boolean formula ψ, where all variables which occur in ψ are ancestors of A in the causal graph of M_V. There are no directed paths from A to any of the variables which occur in ψ. Further, suppose $\langle M, V \cup \{L_A\}\rangle[V] \vdash \psi$. Let $\mathcal{D}$ be a corresponding non-redundant deduction such that all forward-directed inferences by a structural equation are within an indirect proof. Then $\langle M, V \cup \{\neg\psi\}\rangle[V] \vdash \neg L_A$. And there is a corresponding deduction such that all inferential steps by a structural equation are direct, and each forward-directed inferential step to a literal by a structural equation is used in a subproof in the deduction $\mathcal{D}$.

Proof. Let $\langle M, V\rangle$, L_A, and ψ be as assumed in the lemma. Then we know that there is a derivation of ψ from L_A, given $M_V \cup V$. Now, let us take $\neg\psi$

as a premise and the derivation of ψ from L_A, given $M_V \cup V$, as a subproof. By Negation Introduction, we can infer $\neg L_A$. Thus we have obtained a deduction of $\neg L_A$ from $\neg\psi$, given the premises in $M_V \cup V$. Hence, we have shown that $\langle M, V \cup \{\neg\psi\}\rangle[V] \vdash \neg L_A$.

Let us now take a closer look at the deduction of ψ from L_A in the context of $M_V \cup V$. Let D be a variable which is a direct descendant of a variable in ψ such that L_D is an intermediate conclusion in the deduction $\mathcal{D}$ of ψ. L_D is needed to infer ψ in this deduction. By Proposition 1, we know that (i) any deduction of a literal L_D from L_A goes along one or more undirected paths between the variable A and the variable D. Furthermore, by Proposition 1 we know that (ii) intermediate conclusions are needed to infer a literal L_D from L_A concerning the variables intermediate between D and A. Since deduction $\mathcal{D}$ does not contain any forward-directed inferential steps in the main proof, we know that (iii) all intermediate conclusions in this deduction from L_A to L_D are on a directed path from variable D to variable A in the causal graph of M_V.

(i), (ii), and (iii) imply that (iv) deduction $\mathcal{D}$ contains one or more inferential pathways such that each sequence of causal conclusions—in the main proof or a subproof within a proof by cases—goes against the direction of a directed path from D to A in the causal graph of M_V. Hence, all causal inferential steps—outside the subproof of an indirect proof—are backward-directed. (An inferential step is causal iff it uses a structural equation directly, or indirectly in a subproof.) By the proof of Lemma 1, we know that (v) such backward-directed inferential steps are inversions of forward-directed inferential steps such that the latter are needed in the backward-directed inferential steps. If L_D is inferred from L_A in a backward-directed manner, then this inference is based on a forward-directed derivation in which $\neg L_A$ is derived from $\neg L_D$. By (iv) and (v), we have established that (vi) the derivation of L_D from L_A is based on forward-directed derivations, which may be used to construct a derivation of $\neg L_A$ from $\neg L_D$.

So far, we have considered only a single inferential path from L_A to L_D. To infer the formula ψ, more than one such path may be needed, though. Suppose, for example, we have the structural equation $A = B \wedge F$ as equation of A. Further, suppose $M_V \cup V$ is agnostic with respect to both B and F. From A we can then infer both B and F by an indirect proof then. In the next steps, we can exploit the structural equations of B and F for further backward-directed deductions. This gives rise to at least two different inferential paths, both of which start from A. And both inferential paths may be needed in the deduction of ψ.

Now, suppose we have the structural equation $A = B \vee F$. Again, suppose $M_V \cup V$ is agnostic with respect to B and F. Then we can infer $B \vee F$ from A. To exploit this disjunction, we have to start two subproofs: one with the assumption B, another with the assumption F. The two subproofs correspond to two different inferential pathways. Suppose we then infer $L_{B'}$ and $L_{F'}$ in these pathways, respectively. In order to complete the proof by cases, we need to derive $L_{B'} \vee L_{F'}$ by Disjunction Introduction in each subproof.

Finally, note that these observations concerning the derivation of conjunctions and disjunctions may be iterated and combined. Each backward-directed inferential step from a literal may give rise to several inferential pathways. And the structural equation of a literal may have a disjunction of conjunctions on the right-hand side. $A = (B \wedge C) \vee F$ is a case in point. However, the details of the structure of inferential pathways may not concern us here for reasons which will become obvious shortly. But it remains important to note that each backward-directed inferential step starts from a literal, simply because each structural equation has a literal on the left-hand side.

Let $\{D_1, \ldots, D_n\}$ be the set of variables such that each D_i is a direct descendant of a variable which occurs in ψ and some literal L_{D_i} is needed as an intermediate conclusion to infer ψ. For all these literals, claim (vi) can be established in the same way it has been established for L_D above. Hence, (vii) for any literal L_{D_i} needed as an intermediate conclusion in the deduction of ψ such that D_i is a direct descendant of a variable in ψ, claim (vi) holds: $\langle M, V \cup \{\neg L_{D_i}\}\rangle[V] \vdash \neg L_A$ such that there is a corresponding deduction which is forward-directed, where all causal inferential steps have been used in the deduction $\mathcal{D}$.

Since D_i is a direct descendant of a variable which occurs in ψ, there is a structural equation $D_i = \phi_i$ such that a variable occurring in ψ occurs in ϕ_i. Note that, from L_{D_i} either $\neg\phi_i$ or ϕ_i is inferred in the deduction of ψ by an indirect proof. Hence, the deduction $\mathcal{D}$ contains intermediate conclusions equivalent to a formula with the following form: $\nu(S_1\phi_1, \ldots, S_n\phi_n)$, which is a Boolean formula in which the formulas $S_i\phi_i$ are connected by disjunctions and conjunctions but not negated. The components $S_i\phi_i$ make up the right-hand side of the structural equation of a variable D_i which has the properties specified above. S_i is a mere placeholder for a negation symbol which may or may not be present.

Since deduction $\mathcal{D}$ is non-redundant, each formula $S_i\phi_i$ is needed for the deduction of ψ. It holds that $V_M \cup V, \nu(S_1\phi_1, \ldots, S_n\phi_n) \vdash \psi$. Hence, it holds that (viii) $V_M \cup V, \neg\psi \vdash \neg\nu(S_1\phi_1, \ldots, S_n\phi_n)$. (ix) $\neg\nu(S_1\phi_1, \ldots, S_n\phi_n)$ is logically equivalent to a Boolean formula $\nu'(\neg S_1\phi_1, \ldots, \neg S_n\phi_n)$, in which the formulas $\neg S_i\phi_i$ are connected by disjunctions and conjunctions but not negated. Given our explanation of the variables D_i $(i \leq n)$, (vii) implies (x) for all D_i, $M_V \cup V, \neg L_{D_i} \vdash \neg L_A$ such that there is a forward-directed deduction. Further, note that (xi) $L_{D_i} = S_i\phi_i$ since $S_i\phi_i$ is derived from L_{D_i} by an indirect proof. Finally, (viii), (ix), (xi), and (x) imply that that there is a deduction of $\neg L_A$ from $\langle M, V \cup \{\neg\psi\}\rangle[V]$ such that all inferential steps are forward-directed and used in subproofs in the deduction $\mathcal{D}$. This concludes the proof of the lemma.

□

We are now in a position to prove Proposition 5.

Proof. Let $\langle M, V\rangle$ be a causal model which is uninformative on C and E. We prove Proposition 5 by showing that the following implication holds: if there is no direct deduction of E from $\langle M, V\rangle[V][C]$ which has an active path, then there is no indirect deduction of E from $\langle M, V\rangle[V][C]$ which has an active path. To prove this implication, suppose there is no direct deduction of the effect which has an active path. Then suppose, for contradiction, there is an indirect deduction of E from $\langle M, V\rangle[V][C]$ which has an active path.

For the sake of generality, we need to consider combinations of forward-directed and backward-directed deductions. The general form of such a combination may be characterized as follows: we infer from $M_{V\cup\{C\}} \cup V \cup \{C\}$ some sentence ψ such that this sentence concerns one or more variables which are intermediate between C and E in the causal graph $M_{V\cup\{C\}}$. Then we infer $\neg\psi$ from $M_{V\cup\{C\}} \cup V \cup \{C, \neg E\}$. Since the variables of ψ are intermediate between C and E and since E is a descendant of C, we need backward-directed inferences for the latter deduction. The case where $\neg C$ is simply derived from $M_{V\cup\{C\}} \cup V \cup \{C, \neg E\}$ is contained in this consideration as a limiting case: we simply have $\psi \equiv C$ then.

In more formal terms, an indirect proof which combines forward-directed and backward-directed reasoning looks as follows. We show (i) that $M_{V\cup\{C\}} \cup V \cup \{C\}, \neg E \vdash \neg\psi$ such that there is a corresponding deduction which contains one or more backward-directed derivations and which satisfies the condition that each inferential step depends on C. Then we show that (ii) $M_{V\cup\{C\}} \cup V, C \vdash \psi$ such that there is a corresponding deduction which is direct and which satisfies the condition that each inferential step depends on C. These two deductions can be merged so as to obtain a deduction of E from $M_{V\cup\{C\}} \cup V \cup \{C\}$ which has an active path.

However, by Lemma 2 we can infer from (i) that (iii) $M_{V\cup\{C\}} \cup V \cup \{C\}, \psi \vdash E$ such that there is a corresponding deduction which is direct and which satisfies the condition that each inferential step depends on C. Notice that $M_{V\cup\{C\}} \cup V \cup \{C\} \vdash \psi$ and $M_{V\cup\{C\}} \cup V, \psi \vdash E$ imply that $M_{V\cup\{C\}} \cup V \cup \{C\} \vdash E$ by the transitivity of the deduction relation $\vdash$. Moreover, by (ii) and (iii) we know that we can obtain a deduction of E from $M_{V\cup\{C\}} \cup V \cup \{C\}$ by merging two direct deductions both of which satisfy the condition that each inferential step depends on C. The resulting deduction is direct and satisfies the condition that each inferential step depends on C. But this contradicts our initial assumption that there is no direct deduction of the effect E which has an active path, and so we have obtained a contradiction.

Again, for the sake of generality, we need to also consider the case where ψ concerns variables which are descendants of both C and E. In this case, both the deduction of $\neg\psi$ from $\neg E$ and the deduction of ψ from C may be forward-directed. But the combined deduction contains an indirect proof since we still need to assume $\neg E$ in a subproof. Now, we need to distinguish two cases: (i) the deduction of ψ from C contains E as an intermediate conclusion. Then the indirect deduction via ψ would be redundant, and cannot have an active path. (ii) The deduction of ψ from C does not contain E as intermediate conclusion. Then the inferential path from C to E goes along a section of the form $B \rightarrow F \leftarrow D$ in the causal graph of $M_{V\cup\{C\}}$, where F occurs in ψ. But we have shown in the proof of Proposition 2 that this is impossible (see discussion of case (b) in this proof).

Thus we have shown the above implication: if there is no direct deduction of E from $\langle M, V\rangle[V][C]$ which has an active path, then there is no indirect deduction of E from $\langle M, V\rangle[V][C]$ which has an active path. Hence, Proposition 5.

□

3 Chapter 5

Proposition 6. Suppose L_A and L_C are entangled in the causal model $\langle M, V\rangle$. Let $\langle M_N, V_N\rangle$ and $\langle M_E, V_E\rangle$ be as explained in Section 2 of Chapter 5. Then it holds that

(1) $\langle M_N, V_N \cup \{L_A\}\rangle \vdash L_C$ or $\langle M_N, V_N \cup \{L_C\}\rangle \vdash L_A$, or both, and

(2) $\langle M_E, V_E \cup \{L_A\}\rangle \not\vdash L_C$ and $\langle M_E, V_E \cup \{L_C\}\rangle \not\vdash L_A$.

Proof. Suppose L_C, L_A, and the other symbols are understood as explained in the proposition. First, we show claim (2). Claim (1) then follows from claim (2), as we will show below.

Suppose, for contradiction, that $\langle M_E, V_E \cup \{L_A\}\rangle \vdash L_C$ or $\langle M_E, V_E \cup \{L_C\}\rangle \vdash L_A$. Without loss of generality, we assume $\langle M_E, V_E \cup \{L_C\}\rangle \vdash L_A$. (The demonstration for the other assumption is completely analogous.) By Proposition 1, this implies that there is a deduction $\mathcal{D}$ of L_A from $\langle M, V \cup \{L_c\}\rangle$ and an undirected path $\langle C, D_1, \ldots, D_n, A\rangle$ ($n > 0$) of variables such that $\mathcal{D}$ contains an intermediate conclusion for each variable $D_1, \ldots, D_n$. Let us further assume that $\mathcal{D}$ is non-redundant. This assumption will be lifted later on.

Since M_E does, by definition, not contain the structural equations of C and A, there is no directed path from C to A, or vice versa, in the causal graph of M_E. Hence, there is a common effect E of A and C such that (i) deduction $\mathcal{D}$ contains an inferential path from L_C to L_A which goes via E. (ii) This inferential path contains a subsection $\langle L_D, L_E, L_F\rangle$ such that $\langle D \rightarrow E \leftarrow F\rangle$ is an undirected path between D and F in the causal graph of M_E. And (iii) there are no edges among D, E, and F other than (D, E) and (F, E). The variables D and F may or may not coincide with the variables C and A, respectively.

We have seen in the proof of Proposition 2 that (i), (ii), and (iii) imply a contradiction. We have shown this when considering case (b) in this proof.

Let us review here the proof idea. For simplicity, we assume $V_E = \emptyset$. The general case where V_E may not be empty has been considered in the proof of Proposition 2. Suppose (iv) no proof by cases is used to infer L_F. Then no backward-directed inferences are available. We must therefore infer L_F from L_E, which is inferred from L_D. But this is impossible: if we can infer L_E from L_D by the structural equation of E, then the value of E is determined by L_D, and thus independent of the value of F. No inference can be made from L_E to L_F then. If, by contrast, the value of E depends on the value of F in the context of L_D, then we cannot infer the value of E from L_D. Finally, since there are no edges among the variables D, E, and F other than the two edges (D, E) and (F, E), the structural equation of E is the only equation which connects D with E, and F with E.

Let us now assume that (v) deduction $\mathcal{D}$ may contain one or more proofs by cases. Then backward-directed reasoning becomes available with regard to an assumption about E. This may be done in a subproof which starts with the assumption L_E. The corresponding subproof may then contain an inference from L_E and L_D to L_F. However, we need to do a second subproof which starts with an assumption equivalent to $\neg L_E$. If the second subproof starts with a backward-directed inference, then we would have to infer $\neg L_F$ in this subproof. This, however, implies that $\mathcal{D}$ is redundant since $L_F \vee \neg L_F$ is a logical truth in our logic of causal reasoning. But we have assumed that $\mathcal{D}$ is non-redundant.

(vi) All inferences of the second subproof (which starts with an assumption equivalent to $\neg L_E$) must therefore be forward-directed unless we start a second proof by cases. (vii) For such a second proof by cases, we would run into the same problem just observed: there must be a subproof such that all causal inferences in this subproof are forward-directed. Finally, (viii) it must be possible to infer L_A in both subproofs or another literal from which L_A can be inferred. Otherwise, we could establish at most a disjunction of literals by the proof by cases rather than the literal L_A in the deduction $\mathcal{D}$. (vi), (vii), and (viii) imply that there is a directed path from C to A in the causal graph of M_E. Note, however, that this can be ruled out since M_E does not contain the structural equations of C and A by definition. Thus we have obtained a contradiction.

Let us finally lift the assumption that $\mathcal{D}$ is redundant. Since any redundant deduction can be transformed into a non-redundant one by eliminating inferential steps, a non-redundant deduction does not give us additional inferential power over redundant ones. It must therefore also hold for redundant deductions that $\langle M_E, V_E \cup \{L_C\}\rangle \vdash L_A$ implies that there is a directed path from C to A in the causal graph of M_E. Thus have obtained a contradiction for both redundant and non-redundant deductions. Hence, claim (2) of the proposition.

By claim (2) we know that (i) any deduction of L_A from $\langle M, V \cup \{L_C\}\rangle$ does not use an inferential pathway via literals whose variables are a descendant of C or A. Likewise, (ii) any deduction of L_C from $\langle M, V \cup \{L_A\}\rangle$ does not use an inferential pathway via literals whose variables are a descendant of C or A. Since, however, L_C and L_A are by assumption entangled, it holds that $\langle M, V \cup \{L_C\}\rangle \models L_A$ or $\langle M, V \cup \{L_A\}\rangle \models L_C$. By soundness of our deductive system, this implies that $\langle M, V \cup \{L_C\}\rangle \vdash L_A$ or $\langle M, V \cup \{L_A\}\rangle \vdash L_C$. By Proposition 1, this implies (iii) that there is a deduction $\mathcal{D}'$ of L_A from $\langle M, V \cup \{L_C\}\rangle$—or a deduction $\mathcal{D}''$ of L_C from $\langle M, V \cup \{L_A\}\rangle$—such that this deduction goes along one or more undirected paths $\langle A, D_1, \ldots, D_n, C\rangle$ in the causal graph of M.

Without loss of generality, let us assume the first alternative. (The following demonstration is completely analogous for the other alternative.) By (i) and (ii) we know that the variables $D_1, \ldots, D_n$ are not descendants of A and C in the deduction $\mathcal{D}'$. Hence, neither the structural equations in M_E nor the literals in V_E are used in the deductions $\mathcal{D}'$. Claim (iii) therefore implies that the inferential paths which correspond to $\mathcal{D}'$ go along one or more undirected paths in the causal graph of M_N, which represents the inferential relations among variables which are a descendant of C or A in the causal graph of M. In formal terms, $\langle M_N, V_N \cup \{L_C\}\rangle \vdash L_A$. Hence, $\langle M_N, V_N \cup \{L_C\}\rangle \vdash L_A$ or $\langle M_N, V_N \cup \{L_A\}\rangle \vdash L_C$, or both. This concludes the proof of claim (1) of the proposition. Thus we have established both claims of the proposition.

□

Bibliography

Alchourrón, Carlos. E., Peter Gärdenfors, and David Makinson (1985). On the Logic of Theory Change: Partial Meet Contraction Functions and Their Associated Revision Functions. *Journal of Symbolic Logic* 50: 510–30.

Andreas, Holger (2010). Semantic Holism in Scientific Language. *Philosophy of Science* 77(4): 524–43.

—— (2020). *Dynamic Tractable Reasoning: A Modular Approach to Belief Revision*. Cham: Springer.

Andreas, Holger and Lorenzo Casini (2019). Hypothetical Interventions and Belief Changes. *Foundations of Science* 24(4): 681–704.

Andreas, Holger and Mario Günther (2019). On the Ramsey Test Analysis of 'Because'. *Erkenntnis* 84: 1229–62.

—— (2020). Causation in Terms of Production. *Philosophical Studies* 177(6): 1565–91.

—— (2021a). Difference-Making Causation. *Journal of Philosophy* 118(12): 680–701.

—— (2021b). A Ramsey Test Analysis of Causation for Causal Models. *The British Journal for the Philosophy of Science* 72(2): 587–615.

—— (2024a). A Lewisian Regularity Theory. *Philosophical Studies* 181(9): 2145–76.

—— (2024b). A Regularity Theory of Causation. *Pacific Philosophical Quarterly* 105(1): 2–32.

—— (2025a). The Epochetic Analysis of Causation Compared to Counterfactual Accounts. In *PhilSci Archive*, URL `https://philsci-archive.pitt.edu/26137/`.

—— (2025b). The Logic Of Causal Models. In *PhilSci-Archive*, URL `https://philsci-archive.pitt.edu/26213/`.

—— (forthcominga). Actual Causation. *dialectica* .

—— (forthcomingb). Factual Difference-Making. *Australasian Philosophical Review* .

Antoniou, Grigoris (1997). *Nonmonotonic Reasoning*. Cambridge, MA: MIT Press.

Balzer, Wolfgang, C. Ulises Moulines, and Joseph D. Sneed (1987). *An Architectonic for Science. The Structuralist Program*. Dordrecht: D. Reidel Publishing Company.

Barrett, Thomas William and Hans Halvorson (2017). Quine's Conjecture on Many-Sorted Logic. *Synthese* 194(9): 3563–82.

Baumgartner, Michael (2013). A Regularity Theoretic Approach to Actual Causation. *Erkenntnis* 78(1): 85–109.

Baumgartner, Michael and Christoph Falk (2019). Boolean Difference-Making: A Modern Regularity Theory of Causation. *The British Journal for the Philosophy of Science* .

Beckers, Sander (2021). Causal Sufficiency and Actual Causation. *Journal of Philosophical Logic* 50(6): 1341–74.

Beckers, Sander and Joost Vennekens (2018). A Principled Approach to Defining Actual Causation. *Synthese* 195(2): 835–62.

Beebee, Helen (2004). Causing and Nothingness. In *Causation and Counterfactuals*, edited by L. A. Paul, E. J. Hall, and J. Collins, pp. 291–308, Cambridge, MA: MIT Press.

—— (2011). Hume's Two Definitions: The Procedural Interpretation. *Hume Studies* 37(2): 243–74.

Bell, John L. (2023). Infinitary Logic. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Metaphysics Research Lab, Stanford University, Fall 2023 edition.

Berto, Francesco and Daniel Nolan (2023). Hyperintensionality. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Metaphysics Research Lab, Stanford University, Winter 2023 edition.

Black, Max (1956). Why Cannot an Effect Precede its Cause. *Analysis* 16(3): 49–58.

Bradley, Richard (2007). A Defence of the Ramsey Test. *Mind* 116(461): 1–21.

Brand, Myles (1980). Simultaneous Causation. In *Time and Cause*, edited by P. van Inwagen, pp. 137–53, Dordrecht: D. Reidel.

Brewka, Gerhard (1991). Belief Revision in a Framework for Default Reasoning. In *The Logic of Theory Change*, edited by André Fuhrmann and Michael Morreau, pp. 206–22, Berlin: Springer.

Brewka, Gerhard, Jürgen Dix, and Kurt Konolige (1997). *Nonmonotonic Reasoning. An Overview*. Stanford: CSLI Publications.

Briggs, Rachael (2012). Interventionist Counterfactuals. *Philosophical Studies* 160(1): 139–66.

Bromberger, Sylvain (1966). Why Questions. In *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, edited by R. Colodny, pp. 86–111, Pittsburgh: University of Pittsburgh Press.

Carnap, Rudolf (1950). *Logical Foundations of Probability*. Chicago: Chicago University of Chicago Press.

—— (1958). Beobachtungssprache und theoretische Sprache. *Dialectica* 12: 236–48.

Cohen, Jonathan and Craig Callender (2009). A Better Best System Account of Lawhood. *Philosophical Studies* 145(1): 1–34.

Davidson, Donald (1969). The Individuation of Events. In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, pp. 216–234, Dordrecht: D. Reidel Publishing Company.

Dowe, Phil (1996). Backwards Causation and the Direction of Causal Processes. *Mind* 105(418): 227–48.

—— (1997). A Defense of Backwards in Time Causation Models in Quantum Mechanics. *Synthese* 112(2): 233–46.

—— (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dummett, Michael (1954). Can an Effect Precede its Cause? *Aristotelian Society Proceedings Supplement* 28: 27–62.

—— (1964). Bringing About the Past. *Philosophical Review* 73(3): 338–59.

Dwyer, Joseph R. and Martin A. Uman (2014). The Physics of Lightning. *Physics Reports* 534(4): 147–241.

Elga, Adam (2001). Statistical Mechanics and the Asymmetry of Counterfactual Dependence. *Philosophy of Science* 68(3): 313–24.

Enderton, Herbert B. (2001). *A Mathematical Introduction to Logic*. San Diego: Harcourt Academic Press.

Faye, Jan (2021). Backward Causation. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University, Spring 2021 edition.

Fine, Kit (1975). Critical Notice of Lewis, Counterfactuals. *Mind* 84(335): 451–8.

—— (2012). Guide to Ground. In *Metaphysical Grounding*, edited by F. Correia and B. Schnieder, pp. 37–80, Cambridge: Cambridge University Press.

Fischer, Enno (2024). Three Concepts of Actual Causation. *British Journal for the Philosophy of Science* 75(1): 77–98.

Flach, Peter A. (2000). On the Logic of Hypothesis Generation. In *Abduction and Induction*, edited by Peter A. Flach and Antonis C. Kakas, pp. 89–106, Dordrecht: Kluwer.

Friederich, Simon and Peter W. Evans (2019). Retrocausality in Quantum Mechanics. In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Metaphysics Research Lab, Stanford University, summer 2019 edition.

Friedman, Michael (1974). Explanation and Scientific Understanding. *The Journal of Philosophy* 71: 1–19.

Frisch, Mathias (2005). *Inconsistency, Asymmetry and Non Locality: A Philosophical Investigation of Classical Electrodynamics*. Oxford: Oxford University Press.

—— (2014). *Causal Reasoning in Physics*. Cambridge: Cambridge University Press.

Galindo, Alberto and Pedro Pascual (1990). *Quantum Mechanics I*. Berlin: Springer.

Gallow, Dmitri J. (2021). A Model-Invariant Theory of Causation. *The Philosophical Review* 130(1): 45–96.

Gebharter, Alexander, Dennis Graemer, and Frenzis H. Scheffels (2019). Establishing Backward Causation on Empirical Grounds: An Interventionist Approach. *Thought: A Journal of Philosophy* 8(2): 129–38.

Grove, Adam (1988). Two modellings for theory change. *Journal of Philosophical Logic* 17: 157–70.

Gärdenfors, Peter (1978). Conditionals and Changes of Belief. In *The Logic and Epistemology of Scientific Change*, *Acta Philosophica Fennica*, volume 30, edited by I. Niiniluoto and R. Tuomela, pp. 381–404.

—— (1986). Belief Revisions and the Ramsey Test for Conditionals. *The Philosophical Review* 95(1): 81–93.

—— (1988). *Knowledge in Flux*. Cambridge, MA: MIT Press.

Günther, Mario (2022). A Connexive Conditional. *Logos and Episteme* 13(1): 55–63.

Günther, Mario and Caterina Sisti (2022). Ramsey's Conditionals. *Synthese* 200(2): 1–31.

Hall, Ned (2000). Causation and the Price of Transitivity. *The Journal of Philosophy* 97(4): 198.

—— (2004). Two Concepts of Causation. In *Causation and Counterfactuals*, edited by J. Collins, N. Hall, and L. A. Paul, pp. 225–76, Cambridge, MA.: MIT Press.

—— (2007). Structural Equations and Causation. *Philosophical Studies* 132(1): 109–36.

Halpern, Joseph Y. (2000). Axiomatizing Causal Reasoning. *Journal of Artificial Intelligence Research* 12(1): 317–37.

—— (2008). Defaults and Normality in Causal Structures. In *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, edited by G. Brewka and J. Lang, pp. 198–208, Menlo Park, CA: AAAI Press.

—— (2013). From Causal Models to Counterfactual Structures. *Review of Symbolic Logic* 6(2): 305–22.

—— (2015). A Modification of the Halpern-Pearl Definition of Causality. *Proc. 24th International Joint Conference on Artificial Intelligence (IJCAI 2015)* pp. 3022–33.

—— (2016). *Actual Causality*. Cambridge, MA: MIT Press.

Halpern, Joseph Y. and Christopher Hitchcock (2010). Actual Causation and the Art of Modeling. In *Heuristics, Probability, and Causality: a Tribute to Judea Pearl*, edited by R. Dechter, H. Geffner, and J. Y. Halpern, pp. 383–406, London: College Publications.

—— (2015). Graded Causation and Defaults. *British Journal for the Philosophy of Science* 66(2): 413–57.

Halpern, Joseph Y. and Judea Pearl (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *British Journal for the Philosophy of Science* 56(4): 843–87.

Hansson, Sven Ove (1992). In Defense of the Ramsey Test. *The Journal of Philosophy* 89(10): 522–40.

—— (1993). Reversing the Levi identity. *Journal of Philosophical Logic* 22(6).

—— (1999). *A Textbook of Belief Dynamics. Theory Change and Database Updating*. Dordrecht: Kluwer.

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

—— (2002). Physical Causation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 33(4): 717–24.

Hempel, Carl G. (1962). Two Models of Scientific Explanation. In *Frontiers of Science and Philosophy*, edited by R. Colodny, pp. 7–34, Pittsburgh: University of Pittsburgh Press.

—— (1965). *Aspects of Scientific Explanation*. New York: The Free Press.

Hiddleston, Eric (2005). Causal Powers. *The British Journal for the Philosophy of Science* 56(1): 27–59.

Hitchcock, Christopher (2001). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy* 98(6): 273–99.

—— (2007). Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review* 116(4): 495–532.

—— (2009). Structural Equations and Causation: Six Counterexamples. *Philosophical Studies* 144(3): 391–401.

—— (forthcoming). Actual Causation and Factual Difference-Making. *Australasian Philosophical Review* .

Huber, Franz (2013). Structural Equations and Beyond. *Review of Symbolic Logic* 6(4): 709–32.

Huemer, Michael and Ben Kovitz (2003). Causation as Simultaneous and Continuous. *Philosophical Quarterly* 53(213): 556–65.

Hume, David (1739/2001). *A Treatise of Human Nature*. Oxford: Oxford University Press. Edited by D. Norton and M. Norton.

Husserl, Edmund (1913/1989). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*. Dordrecht: Kluwer. Translation of Husserl's *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie* by R. Rojcewicz and A. Schuwer.

Kant, Immanuel (1781/1998). *Critique of Pure Reason*. Cambridge: Cambridge University Press. Translation of Kant's *Kritik der reinen Vernunft* by Paul Guyer and Allen W. Wood.

Kim, Jaegwon (1969). Events and Their Descriptions: Some Considerations. In *Essays in Honor of Carl G. Hempel*, edited by Nicholas Rescher, pp. 198–215, Dordrecht: D. Reidel Publishing Company.

Kitcher, Philip (1989). Explanatory Unification and the Causal Structure of the World. In *Scientific Explanation*, edited by Philip Kitcher and Wesley Salmon, pp. 410–505, Minneapolis: University of Minnesota Press.

Kutach, Douglas (2013). *Causation and Its Basis in Fundamental Physics*. Oxford: Oxford University Press.

Leibniz, Gottfried W. (1686/1998). Principles of Nature and Grace. In *Philosophical Texts*, pp. 258–66, Oxford: Oxford University Press. Translation of Leibniz's *Principes de la nature et de la Grâce fondés en raison* by R. S. Woolhouse and R. Francks.

Leitgeb, Hannes (2010). Ramsey Test Without Triviality. *Notre Dame Journal of Formal Logic* 51(1): 21–54.

Lewis, David (1973a). Causation. *The Journal of Philosophy* 70(17): 556–67.

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

—— (1979). Counterfactual Dependence and Time's Arrow. *Noûs* 13(4): 455–76.

—— (1986a). Events. In *Philosophical Papers, Volume II*, pp. 241–69, Oxford: Oxford University Press.

—— (1986b). Postscripts to "Causation". In *Philosophical Papers. Volume II*, pp. 172–213, Oxford: Oxford University Press.

—— (2000). Causation as Influence. *The Journal of Philosophy* 97(4): 182–97.

Loewer, Barry (2007). Counterfactuals and the Second Law. In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, edited by H. Price and R. Corry, pp. 293–326, Oxford: Oxford University Press.

Lutgens, Frederick K., Edward J. Tarbuck, and Dennis Tasa (2013). *The Atmosphere: an Introduction to Meteorology*. Boston: Pearson, 12th edition.

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly* 2(4): 245–64.

—— (1980). *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press.

May, Michael and Gerd Graßhoff (2001). Causal Regularities. In *Current Issues in Causation*, edited by W. Spohn, M. Ledwig, and M. Esfeld, pp. 85–114, Paderborn: Mentis.

McDermott, Michael (1995). Redundant Causation. *The British Journal for the Philosophy of Science* 46(4): 523–44.

McDonald, Jennifer (2020). Strong Proportionality and Causal Claims. In *PhilSci-Archive*, URL `https://philsci-archive.pitt.edu/16809/`.

McGrath, Sarah (2005). Causation By Omission: A Dilemma. *Philosophical Studies* 123(1-2): 125–48.

Meheus, Joke, Christian Straßer, and Peter Verdée (2013). Which Style of Reasoning to Choose in the Face of Conflicting Information? *Journal of Logic and Computation* 26(1): 361–80.

Menzies, Peter and Huw Price (1993). Causation as a Secondary Quality. *British Journal for the Philosophy of Science* 44(2): 187–203.

Mill, John Stuart (1843/2011). *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence, and the Methods of Scientific Investigation*. Longmans, Green, Reader, and Dyer.

Noordhof, Paul (2020). *A Variety of Causes*. New York, NY: Oxford University Press.

Papineau, David (1992). Can We Reduce Causal Direction to Probabilities? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association* 1992: 238–52.

Paul, L. A. and Ned Hall (2013). *Causation: A User's Guide*. Oxford: Oxford University Press.

Paul, Laurie A. (2000). Aspect Causation. *Journal of Philosophy* 97(4): 235–56.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 1st edition.

—— (2009). *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press, 2nd edition.

Peirce, Charles Sanders (1931). *Collected Papers of Charles Sanders Peirce*. Cambridge, MA: Harvard University Press. Edited by C. Hartshorne and P. Weiss and A. Burks.

Poincaré, Henri (1902/1952). *Science and Hypothesis*. New York: Dover Publications, 2nd edition. Translation of Poincaré's *La Science et l'Hypothèse* by G. B. Halsted.

Price, Huw (1992). Agency and Causal Asymmetry. *Mind* 101(403): 501–20.

—— (1996). *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time*. Oxford: Oxford University Press.

—— (2017). Causation, Intervention and Agency–Woodward on Menzies and Price. In *Making a Difference*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, pp. 73–98, Oxford: Oxford University Press.

—— (forthcoming). The Practical Arrow. *Australasian Philosophical Review* .

Putnam, Hilary (1980). Models and Reality. *Journal of Symbolic Logic* 45(3): 464–82.

—— (1981). *Reason, Truth, and History*. Cambridge: Cambridge University Press.

—— (1992). *Realism with a Human Face*. Cambridge, MA: Harvard University Press.

Quine, Willard V. O. (1961). Two Dogmas of Empiricism. In *From a Logical Point of View*, pp. 20–46, Cambridge, MA: Harvard University Press, 2nd edition.

Rakov, Vladimir A. and Martin A. Uman (2003). *Lightning: Physics and Effects*. Cambridge: Cambridge University Press.

Ramsey, Frank Plumpton (1931a). General Propositions and Causality. In *The Foundations of Mathematics and Other Logical Essays*, edited by R. B. Braithwaite, pp. 237–55, London & New York: Routledge.

—— (1931b). Theories. In *The Foundations of Mathematics and Other Logical Essays*, edited by R. B. Braithwaite, pp. 212–36, London & New York: Routledge.

Reichenbach, Hans (1956). *The Direction of Time*. Los Angeles: University of California Press.

Richards, Thomas J. (1965). Hume's Two Definitions of 'Cause'. *Philosophical Quarterly* 15(60): 247–53.

Robinson, J. A. (1962). Hume's Two Definitions of "Cause". *Philosophical Quarterly* 12(47): 162–71.

Rott, Hans (1986). Ifs, Though, and Because. *Erkenntnis* 25(3): 345–70.

—— (2011). Reapproaching Ramsey: Conditionals and Iterated Belief Change in the Spirit of AGM. *Journal of Philosophical Logic* 40: 155–91.

Russell, Bertrand (1913). On the Notion of Cause. *Proceedings of the Aristotelian Society* 13: 1–26.

—— (1918/2010). *The Philosophy of Logical Atomism*. Routledge Classics, London: Routledge.

—— (1921/2009). Psychological and Physical Causal Laws. In *The Basic Writings of Bertrand Russell*, edited by R. E. Egner and L. E. Denonn, pp. 287–95, Routledge Classics, London & New York: Routledge.

Sartorio, Carolina (2005). Causes as Difference-Makers. *Philosophical Studies* 123(1): 71–96.

Schaffer, Jonathan (2000). Trumping Preemption. *Journal of Philosophy* 97(4): 165–81.

Schnieder, Benjamin (2014). Bolzano on Causation and Grounding. *Journal of the History of Philosophy* 52(2): 309–37.

Schurz, Gerhard (2008). Patterns of Abduction. *Synthese* 164: 201–34.

Schurz, Gerhard and Alexander Gebharter (2016). Causality as a Theoretical Concept: Explanatory Warrant and Empirical Content of the Theory of Causal Nets. *Synthese* 193(4): 1073–103.

Sneed, Joseph D. (1979). *The Logical Structure of Mathematical Physics*. Dordrecht: D. Reidel Publishing Company, 2nd edition.

Spirtes, Peter, Clark Glymour, and Richard N. Scheines (1993). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press.

Spohn, Wolfgang (1988). Ordinal Conditional Functions: A Dynamic Theory of Epistemic States. In *Causation in Decision, Belief Change, and Statistics II*, pp. 105–34, Dordrecht: Kluwer.

—— (2006). Causation: An Alternative. *British Journal for the Philosophy of Science* 57(1): 93–119.

—— (2012). *The Laws of Belief: Ranking Theory and its Philosophical Applications*. Oxford: Oxford University Press.

Stalnaker, Robert C. (1968). A Theory of Conditionals. In *Studies in Logical Theory (American Philosophical Quarterly Monograph Series)*, edited by N. Rescher, pp. 98–112, 2, Oxford: Blackwell.

Suppes, Patrick (1957). *Introduction to Logic*. Princeton: Von Nostrand.

Taylor, Richard (1966). *Action and Purpose*. Englewood Cliffs, NJ: Humanities Press.

van Ditmarsch, Hans, Wiebe van der Hoek, and Barteld Kooi (2008). *Dynamic Epistemic Logic*. Berlin: Springer.

van Fraassen, Bas (1980). *The Scientific Image*. Oxford: Oxford University Press.

Weslake, Brad (forthcoming). Commentary on Factual Difference-Making. *Australasian Philosophical Review* .

Wilson, Alastair (2018). Metaphysical Causation. *Noûs* 52(4): 723–51.

Wittgenstein, Ludwig (1953). *Philosophical Investigations*. Oxford: Blackwell. Translation of Wittgenstein's *Philosophische Untersuchungen* by G. E. M. Anscombe.

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

Wysocki, Tomasz (forthcoming). No Causation for the Unsettled. *Australasian Philosophical Review* .

Zach, Richard (2021). *Sets, Logic, Computation: An Open Introduction to Metalogic*. OpenLogicProject, URL `http://builds.openlogicproject.org`.

Index of Names

Hume, 4–5, 303–304

Kant, 151–152, 303–304

Index of Subjects

abductive inference, 172–179
absence, 15–16
absence rule, 111–114
active path, 41–52

bilking argument, 244–251

causal explanation, 216–219
causal graph, 38–40
causal model, 27–38
causation
- backward, 234–287
- concept of, 4–7
- conjunctive, 55–57
- non-transitivity of, 76–90
- simultaneous, 214–233
- spurious, 164–213
- temporal asymmetry of, *see* Humean convention

cause, 44–46, 115–116, 160–161, 204–205, 218–219, 296–298
collaboration, 100–102
common cause, 164–195, 251–256
conjunctive fork, 251–253
convention, 148–151
counterfactual, 260–274
- backtracking, 260–265

counterfactual approach, 21–24, 260–274

default, 111–115
deviancy, 111–116

entanglement, 96–100
epochetic approach, 2–4
epochetic conditional, 2–3, 8–9, 157–159, 327–329
events, 15–20

grounding, 20, 229

Humean convention, 12–15, 145–152, 160–161, 165–166, 216–219, 259–260, 284–287

inferential network, 46–52
inferential path, 40–52
intervention, 34–35, 235–246
INUS condition, 210–211, 277–278
isomorphism, 92–93, 108–110

law, 154–157
- ceteris paribus, 156
- default, 111–115, 156–157

- non-redundant, 171–172, 179–180, 196–203
- redundant, 171–172, 179–180, 196–203

law of nature, 169–170, 207
- best system account of, 207

neuron diagram, 53–55
normality, 111–116

omission, 128–131
overdetermination, 58–59, 136–140

preemption, 46–52, 63–65, 136–140
- early, 59–63
- late, 65–67
- trumping, 68–71

prevention, 71–73
- bogus, 117–118
- bogus double, 120
- double, 73–75
- extended bogus, 118–120
- extended double, 79–82
- isomorphic modified extended double, 126–128
- modified extended double, 124–125

Ramsey Test, 8–9, 157–159, 327–329
reason, 2–3, 8–9, 159–160

Sartorio's principle, 90–93
short circuit, 77–79
- extended, 123–124

structural equation, 28–31
subcause, 94–96
switch
- basic, 86–88
- realistic, 88–90, 131–134
- simple, 83–86

unification, 169–170, 199–207