The Emergence of Causation

by

Jeffrey Dmitri Gallow

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in The University of Michigan
2014

Doctoral Committee:

Professor James M. Joyce, Chair
Professor Gordon Belot
Professor Allan F. Gibbard
Associate Professor Christopher L. House
Professor Brian J. Weatherson

In memory of my father,
James F. Gallow

Acknowledgements

I have now officially joined the ranks of the over-educated. How could this have happened? Where did things go wrong? Who is to blame? Undoubtedly, some of the blame must lie with me, but I did not act alone. There were those who goaded me on at each step of the way, and they each deserve their fair share of the blame.

I grew up in Snellville, Georgia, a small town on the outskirts of Atlanta; I spent the first 13 years of my education in Catholic school. I was 16 before I had a friend who wasn't Catholic. I was about the same age when I first had a friend who wasn't politically conservative. During elementary school, I was prohibited from checking some books out of the library because they were considered too advanced. I was frequently banished to the hallway after disagreeing with my religion teachers about women in the priesthood or the morality of homosexuality. This was not an environment that was especially welcoming or encouraging of independent critical thought. Conditions were prime for me to develop into an academically successful but altogether boring and intellectually spineless person. To the extent that that didn't happen, it is due to the fact that, when I was young, my father's illness forced him into early retirement. This meant that he ended up taking over all of the traditionally maternal duties; in particular, it meant that he would drop me off at school in the morning and pick me up in the evening. During the car rides home, Dad would ask me what I had learned that day in school. When I told him, he would do what no teacher ever had: he would present me with alternate viewpoints, competing explanations, objections; he would ask me to decide which viewpoint I thought was correct, to defend the stance I had taken against objections, to say why the alternatives were wrong. Like many fathers and sons, Dad and I didn't have much to bond over; we didn't have much in common or much that we agreed on; but in those car rides, we found some commonality in disagreement, in a shared appreciation of argumentation. Those car rides, more than anything that followed, are to blame for what I have become.

Prior to college, good teachers were rare, but that made the exceptions all the more

important. In their own ways, Ms. Jenner, Mrs. Frapwell, and Mrs. Koon each provided some desperately needed encouragement and direction. Dr. Cynthia Schafer remains one of the best teachers that I've ever had. As an undergraduate, my decision to major and pursue graduate study in philosophy was largely influenced by the instruction of John Devlin, Greg Fitch, Patricia Kitcher, Bernard Kobes, Carol Rovane, and Achille Varzi, among others. About nine years ago now, I was enrolled in David Albert's philosophy of science course as an undergraduate. With characteristic verve, Professor Albert left me puzzled about some matters that would previously have struck me as mundane. Those were the matters of causation and chance. I have yet to emerge from that state of aporia, and the topic of this dissertation is fairly well explained by my time in Professor Albert's lectures. These teachers and professors have each earned, in their own ways, some share of the blame for the writing of this dissertation.

Were it not for the almost daily philosophical conversations I've had with my fellow graduate students over the past six years, I would know far, far less about the field of philosophy than I do today. I must blame every graduate student at Michigan for the invaluable lessons they have given me over the years, but some have done enough damage to warrant being singled out. Daniel Drucker has had a consistently bad influence; during the year that we lived together, Daniel lured me into hour-long philosophical debates on an almost nightly basis. I must also blame Chloe Armstrong, Aaron Bronfman, Billy Dunaway, Lina Jansson, Zoë Johnson-King, Dustin Locke, Neil Mehta, Brian Parkhurst, Cat Saint Croix, Chip Sebens, Jon Shaheen, Daniel Singer, Rohan Sud, and Robin Zheng for many instructive conversations.

I have learned an enormous amount from all of the faculty here at Michigan, but special blame is due to David Baker—who read and gave helpful comments on my prospectus, chapter drafts, and my writing sample for the job market—Maria Lasonen-Aarnio—who read and influenced drafts of my work on epistemology—David Manley—who is always eager to talk philosophy, and whose work in metaphysics has shaped my own thinking in myriad ways—Sarah Moss—from whom, along with Eric Swanson, I have learned just about everything I know about the philosophy of language, who was always willing to read drafts of material, and whose comments have frequently set the agenda for future revisions—and Laura Ruetsche—who is most to blame for any semblance of understanding of statistical mechanics and quantum mechanics I possess.

An inordinate amount of blame accrues to those on my committee, who not only incited me to write the dissertation by signing off on the prospectus, but also provoked

revisions and improvements with their comments. Beyond their devious machinations to educate me beyond any reasonable level, and to keep me in the philosophical profession for a little while longer, at least, Jim, Gordon, Allan, and Brian have each, with the influence of their own groundbreaking, agenda-setting work, indoctrinated me with an undue sense of optimism about the possibility of philosophical progress, and a zeal to do philosophical questions the justice they deserve.

Besides those already mentioned, I must blame Katie Elliot, Clark Glymour, Daniel Hausman, Christopher Hitchcock, David Kaplan, Eric Lormand, Calvin Normore, Peter Railton, Patrick Shirreff, Sheldon Smith, and audiences at the University of Michigan, UCLA, Illinois Wesleyan University, Leeds University, Koç University, and Lingnan University for helpful feedback on this material.

·

I arrived in Ann Arbor six years ago, and I have spent an undue amount of the time since hating it. As often happens with the things I hate, I have begun to miss it terribly now that I am leaving it behind. These have been at once the most desperate, anxiety-ridden, lonely, inspiring, exciting, warm, and meaningful years of my life, and I find myself unexpectedly saddened to leave them behind—to leave behind the city that I never wanted to be, but ended up being all the same, my home. A handful of amazing people—Armstrong, Bryson, Dunaway, Drucker, Edmonds, Kornberg, Jansson, Johnson-King, Parkhurst, Saint Croix, Singer, Sud, Wassel, and Webster, to name a few—have made the years here more inspiring, exciting, warm, and meaningful than they would otherwise have been. I cannot blame them enough for being the incredible people that they are; I would not have made it through graduate school without their companionship. Rohan: you got me through a tough time in my life—thank you so much. Zoëtrope: thanks for the controlled chaos, the jokes, droodles, and riddles. Here's one more for you—good luck: *orc sdre m ;sfmh'o so ;rusmoo so .nomkhfh suh orc*. Sabeen: it's been just about ten years since that morning in Bucharest when Dr. Pickus told me that I should talk to you, and it remains the best advice I've ever received. Most everything good in my life has come as a result of meeting you. I marshmallow you. Mom: you are the sole reason that I'm anywhere even close to emotionally well-adjusted; some of the best times of my life thus far were the silly, private jokes we shared; I love you.

Preface

Many contemporary philosophers are drawn to a doctrine of *causal fundamentalism*. This doctrine has its epistemic and its metaphysical tenets. The epistemic tenet is that causation is to be investigated entirely, or primarily, by looking to the ground-floor level of reality. For some—mostly philosophers of science—this means looking to the theories of fundamental physics. Whatever causation is, it is the kind of thing which will be revealed through careful study of the fundamental physical state of the world and the fundamental physical laws of nature. For others—mostly metaphysicians—this means looking to fundamental metaphysics. Whatever causation is, it is the kind of thing which will be revealed through constructing fundamental metaphysical theories of causation and weighing their theoretical virtues against one another. According to the causal fundamentalist, causation is emphatically *not* to be investigated through what has come to be known as 'conceptual analysis'. It is not be investigated by constructing theories of causation, and weighing them against one another in terms of how well they capture our pre-theoretic causal judgments.

The metaphysical tenet of causal fundamentalism is that, once the causal structure of the fundamental level of reality is settled—once, that is, we know which fundamental physical events cause which other fundamental physical events—*all* of the world's causal structure is settled. This could either be because the fundamental causal structure is the only causal structure that there is to settle, or it could be because all other causal structure that there is in some sense *reduces* to that found at the fundamental physical level.

The first part of the dissertation—consisting of chapters 1 and 2—disputes both tenets of causal fundamentalism. Chapter 1 argues against the methodology, advocated by Dowe (2000), among others, of attempting to locate causation within the theories of fundamental physics without reference to any of our pre-theoretic causal judgments—without, that is, engaging in conceptual analysis. (The arguments of that chapter generalize straightforwardly to those who wish to eschew conceptual analysis in favor of investigating the "fundamental metaphysics of the causal relation"[1] through *a priori* insight and considerations of theoretical simplicity.) Chapter 2 argues that the world comes equipped with novel and irreducible causal structure at higher levels of description. Settling the world's fundamental causal structure does not suffice to settle all of the world's causal structure. Accepting this thesis does not commit us to the view that there are irreducible higher-level laws, or irreducible higher-level properties. We can accept this thesis while consistently maintaining that, once the fundamental physical laws of nature and the fundamental physical state of the world are settled, *all* causal facts, at every level of description, are settled as well.

The second part of the dissertation—consisting of chapters 3, 4, and 5—takes up the task of constructing an anti-fundamentalist theory of causation. On this theory, the causal relata are the events of parts of the world acquiring or retaining certain properties. (Throughout, I'll call the causal relata 'events', but I mean to remain neutral on the question of whether the causal relata are events or facts or perhaps something else; my theory excludes certain 'thick' theories of events according to which no distinct events share the same spatiotemporal boundaries,[2] but beyond this, I can be ecumenical.) In order for two events such as these to be causally related, one must counterfactually depend, in at least one way, upon the other, or else there must be a certain kind of chain of counterfactual dependence leading from the one to the other. This theory is broadly in the spirit of Lewis's 1973 counterfactual account of causation. Yet, as chapter 5 demonstrates, it is able to escape the primary counterexamples which led even Lewis to eventually abandon that theory. It is the groundwork laid in chapters 3 and 4 which allows the theory to elude these counterexamples.

The first bit of that groundwork comes in chapter 3, with the introduction of what I call *structural determination relations*. These are the relations which are represented in *structural equations models*, which have become increasingly prevalent in philosophical theorizing about causation. It remains opaque, however, what exactly it takes for one of these models to be *correct*—what it takes, that is, for one of these models to represent a network of structural determination relations correctly. To the extent that this question has been addressed, the answer most have found congenial is that these models are correct if and only if a certain family of causal counterfactual conditionals are true. This answer forsakes any hope of offering a reductive analysis of causal counterfactuals in terms of relations of structural determination. More-

[1] Sider (2011, p. 16, fn 7)

[2] See, *e.g.*, Quine (1986)

over, if we had hoped to offer a reductive analysis of causation in terms of causal counterfactuals, then this answer forsakes any hope of offering a reductive analysis of causation in terms of relations of structural determination. In chapter 3, I note that there are other, deeper, problems with this answer as well. In particular, it cannot explain why relations of structural determination have the properties that they are standardly assumed to have—in particular, it cannot explain why relations of structural determination are *independently manipulable*. I therefore propose an alternate understanding of structural determination which 1) allows us to give a reductive analysis of causal counterfactual dependence—one which neatly solves a problem case for the Lewis/Stalnaker 'closest possible world' semantics; and 2) explains why relations of structural determination are independently manipulable.

Because causal counterfactuals are evaluated relative to networks of structural determination, and because causal counterfactuals provide the truth conditions for singular causal claims, these networks of structural determination represent the pathways along which singular causal influence propagates. On the account of structural determination I outline and defend in chapter 3, merely because the *fundamental* state of some part of the world structurally determines the *fundamental* state of some other part of the world, this does not mean that the higher-level properties of the former part of the world structurally determine the higher-level properties of the latter part of the world. For this reason, settling the world's fundamental causal structure does not suffice to settle all the world's causal structure.

In providing an account of causation which soars to such ontic heights, we risk losing sight of some more minute details of the world which end up making differences to the question of which events singularly caused with other events. In chapter 4, this risk becomes manifest. There, I review Hall (2007)'s objection that we may have two systems which—for all I or anyone else has said about what it takes for a structural equations model to be correct—are correctly modeled by the very same structural equations model. Nevertheless, these two systems differ from one another causally. In one, the event C caused the event E; whereas, in the other, the event corresponding to C *didn't* cause the event corresponding to E. This means that, given all I or anyone else has said about what it takes for a structural equations model to be correct, a correct structural equations model radically *underdetermines* the world's causal structure. Many have hoped that this problem could be dealt with by including information about which states of the world are *default*, and which are noteworthy *deviations* therefrom. In chapter 4, I show that these hopes are in vain. For there are pairs of systems which—for all I or anyone else has said—may be correctly modeled by a single structural equations model; moreover, the corresponding events in each system are all equally deviant; yet, in one system, C caused E, whereas, in the other system, the event corresponding to C didn't cause the event corresponding to E. After a discussion of the various options for dealing with this problem, I ultimately settle upon the solution that, in order to reveal singular causal information, all of the structural determination relations in a structural equations model must be *autonomous*, in a sense made precise in chapter 4.

With this groundwork laid, I proceed, in chapter 5, to demonstrate that relatively minor alterations to Lewis's 1973 account of causation as the ancestral of counterfactual dependence suffice to avoid the numerous counterexamples to that view. The primary differences between Lewis's account and my own are four-fold: firstly, whereas, for Lewis, the relevant counterfactual to consider is 'if this event had not occurred, then some other event would not have occurred either', on my account, the relevant counterfactuals are roughly of the form 'had this property of the world not been what it was (or had not changed when it did), then this other property of the world would not be what it was (or would not have changed when it did)'. Because these counterfactuals do not concern the occurrence or nonoccurrence of events, we need not say anything about the modal profiles of events in order to say whether they are causally related or not.[3] Secondly, Lewis evaluated his causal counterfactuals using the 'closest possible world' semantics of Stalnaker (1968), Lewis (1973a, 1979b). On my account, the counterfactuals are evaluated with reference to networks of structural determination. One consequence of this difference is that, whereas Lewis could only consider 'global' counterfactuals, my account allows us to consider what I call 'local' counterfactuals as well—in evaluating a local counterfactual, one looks only at the parts of the world that immediately structurally determine another part of the world, and ignores any determination relations that may exist between those determiners. So, if whether the window shatters is determined by whether Suzy throws and whether Billy throws, and whether Billy throws is determined by whether Suzy throws (he will throw iff she doesn't), then while the global counterfactual 'if Suzy hadn't thrown, the window wouldn't have shattered' is false, the *local* counterfactual 'if Suzy hadn't thrown, then window wouldn't have shattered' is true—because, in evaluating this counterfactual, we hold Billy's not throwing fixed. Thirdly, for Lewis, it did not mat-

[3] *Cf.* Paul (2000, 2004) and Paul & Hall (2013, p. 237 ff.)

ter what kinds of events were being considered—not whether they involved interesting deviations from normality, nor whether they involved a property of the world *changing* as opposed to remaining constant. On my account, these differences matter. The kinds of counterfactuals which must be considered differ depending upon whether we are talking about changes in the state of the world; and whether a chain of counterfactual dependence is sufficient for causation depends, in part, upon whether the events appearing in the chain are deviations from normality. Fourthly and relatedly, on Lewis's account, *whenever* there is a chain of counterfactual dependence running from C to E, this is sufficient for C's causing E. On my account, this is only sufficient if three additional conditions are met: 1) the chain doesn't leave out any important part of the causal process leading from C to E, in a sense to be made precise in chapter 5; 2) every event appearing in the chain is an interesting or noteworthy deviation from inertial conditions; and 3) the counterfactuals in the chain are all *interlocking*, in the sense that the conditions which make the consequent of a counterfactual in the chain true are the conditions considered in evaluating the antecedent of its successor in the chain. So emended, I show, throughout chapter 5, that the account can deal with cases of early and late preemption, it can accommodate the violations of transitivity, preemptive prevention, short circuits, symmetric overdetermination, preemptive double prevention, and trumping.

Table of Contents

List of Figures

Figure

CHAPTER 1

THE PROJECT OF THEORIZING ABOUT CAUSATION

1.1 Introduction

THE philosophy of causality is rife with reflection on and disagreement about its methodological foundations. For the better part of forty years, at least, philosophers working in the wake of David LEWIS have investigated causality by building theories around our *characteristic causal judgments*. Given a scenario described non-causally, do we judge the English sentence '*C* caused *E*' to be true or false (assertible or unassertible)? Roughly, a good account is one which aligns with our characteristic causal judgments, and a bad account is one which departs from our characteristic causal judgments.[1] In sharp contrast, the intellectual descendants of Bertrand RUSSELL and Hans REICHENBACH—most notably, Wesley SALMON and Phil DOWE—have constructed accounts of causality which attempt to identify causality with some feature of our best fundamental physical theories—in the case of DOWE and SALMON,

[1] I don't mean to suggest that this methodology was in any way *original* with LEWIS. Indeed, as will become clear below, I believe that the methodology is implicit even in those, like DOWE (2000), who explicitly deny that they utilize this methodology.

causality is identified with, roughly, the transmission or retention of a conserved quantity. According to Dowe, the ground level data for his theory of causality is given by our fundamental physical theories. On Dowe's view, just as our physical intuitions must give way to mature theories of physics, so too must our causal intuitions.

Dowe (2000) dedicates his introductory chapter to defending his method of theorizing about causation. He speculates, I believe correctly, that many disputes between philosophers of causality boil down to metaphilosophical disputes about how to properly investigate causality—disputes about, *e.g.*, what constitutes evidence in favor of a theory of causality. In this chapter, I will follow Dowe's lead, and attempt to explain how I will be approaching the task of providing a philosophical theory of causality, and why I will be approaching it in that way. Readers who are already broadly sympathetic to the Lewisian methodology should feel free to skip ahead to chapter 2.

I will begin by arguing against Dowe (2000)'s view that a theory of what causation is in the world should be blind to our characteristic causal judgments. I expect, however, that merely arguing against this position will be inadequate to move those philosophers who find it mysterious that mere introspection could reveal substantive truths about the nature of causation. More effective, I believe, than arguing against positions like Dowe's is offering a rival understanding of the epistemology underlying the methodology I am offering. My approach will be to argue that, while c's being causally related to e has nothing to do with our concepts, and everything to do with the way that the world actually is, the fact that the causal relation is the thing which is picked out by our concept of causation has much to do with the contours of that concept. And that means that we can use our implicit knowledge of the applicability conditions of the concept of causation to investigate the mind-independent relation which we call 'causation'.

1.2 Empirical and Conceptual Analyses

Dowe (2000) distinguishes two tasks which a philosopher of causation might pursue: first, "to elucidate our normal concept of causation,"[2] and second, "to discover what causation is in the objective world."[3] The same distinction is drawn by Psillos (2009), who separates "a theory of the meaning of causal statements" from

[2] Dowe (2000, p. 1)
[3] Dowe (2000, p. 1)

“a theory of what causation is in the world [or] a theory about the worldly constituents of causation”[4] Dowe (2000) calls the first task that of providing a ‘conceptual analysis’ of causation and the second, that of providing an ‘empirical analysis’ of causation. Conceptual analysis is, according to Dowe, *a priori*, and the facts it discovers are necessary. Empirical analysis, in contrast, must be *a posteriori*, and the facts that it turns up may be only contingently true (indeed, Dowe believes that his account of causality would be refuted if certain solutions to the Einstein field equations were actual).[5] According to Dowe, each of these two tasks is legitimate, but they must not be confused with one another. In particular—and this is where Dowe and I part company—discovering something about our concept of causation does not tell us anything about what causation is in the world. Similar thoughts show up in Sider (2011), who chastises the causation literature for failing

> to make up its mind whether it is about fundamental metaphysics or conceptual analysis...If the project is conceptual analysis, then heavy reliance on thought experiments is appropriate...If, on the other hand, the project is to investigate the fundamental metaphysics of the causal relation, then...heavy reliance on thought experiments must be abandoned... (Sider, 2011, p. 16, fn. 7)

While there is an interesting epistemological point here to be ferreted out, Dowe’s distinction between *the concept* of causation and causation *as it is in the world* is largely a sideshow. There is surely a distinction to be drawn between the concept CAUSATION, expressed by the English verb ‘to cause’, and the relation ⟨CAUSATION⟩ out in the world. (To emphasize this difference, I will use small caps to denote concepts and the angle bracketed ‘⟨CONCEPT⟩’ to denote the thing out in the world denoted by CONCEPT.) However, simply because these two are distinct does not mean that a theory of one does not harbor substantive commitments about the other. There is, after all, just as surely a distinction to be drawn between 1 and 2.

1. ⟨ The cat is on the mat ⟩ is true.[6]

2. The cat is on the mat.

[4] Psillos (2009, p. 133)

[5] See Dowe (2000, p. 97).

[6] I use ‘$\langle p \rangle$’ to denote the proposition picked out by the sentence ‘p’ in a given context of use.

1 attributes a property to a proposition; whereas 2 attributes a property to a cat.[7] Nevertheless, it is a truism—just about the only uncontroversial truism about truth—that 1 holds iff 2 does. Thus, we should be able to infer 2 from 1. But to say this is just to say that we can infer facts about the world—*e.g.*, that the cat is on the mat—from facts about our language or our thoughts—*e.g.*, that the English sentence 'the cat is on the mat' picks out a true proposition. And there is precisely the same kind of tight connection between the *concept* CAUSATION and the *relation* ⟨CAUSATION⟩. Specifically, 3 holds iff 4 does.

3. CAUSATION applies to c and e

4. ⟨CAUSATION⟩ relates c and e

And this tells us that we can infer facts about which events (or facts, or whatever) are causally related out in the world from facts about which events (or whatever) our concept of causation applies to.

I expect that authors like Dowe and Psillos would accept that, in some good sense of the world 'concept', the concept CAUSATION applies to c and e iff the relation ⟨CAUSATION⟩ relates c and e. What they ought to challenge is the idea that, so understood, our concept of causation can be fruitfully investigated from the armchair—that is, what they ought to challenge is the suggestion that our characteristic causal judgments tell us anything about the concept CAUSATION—in the sense that 3 holds iff 4 does. Dowe, for instance, likens the concept of CAUSATION to the concept of ENERGY (p. 7). Our characteristic energy judgments might tell us that Zoë is more energetic than Daniel. However, it would be wrong to infer anything about Zoë and Daniel's respective masses from this pre-theoretic judgment, together with the empirical discovery that energy is proportional to mass.

This much is surely correct. Dowe goes on to allege that ENERGY had a pre-theoretic meaning, characterized by judgments like 'Zoë is more energetic than Daniel', before it was given a precise physical definition. From there, he argues by analogy that 'causation' may be given a precise physical definition, and that for that reason we shouldn't defer to our pre-theoretic causation judgments any more than we defer to our pre-theoretic energy judgments. He writes that

> We can say that application of the scientific method of theorizing and experimentation produced an 'empirical analysis' of energy. In the same

[7] I'm just assuming here that pro-sentential theories of truth are incorrect.

> way, science may reasonably be expected to throw light on the language-independent entity called 'causation.'[8]

I find these remarks confusing. The physicist's use of the word 'energy' in sentences like 'Energy is proportional to mass' appears to be polysemous with my use of the word 'energy' in an utterance of 5.[9]

5. Zoë has more energy than Daniel.

Simply because Daniel weighs more than Zoë—and therefore, has more physical energy—this does not make my utterance of 5 false. We sometimes use the word 'bank' to mean a financial institution; and we sometimes use the word 'bank' to mean the building in which a financial institution operates.[10] So, in one sense of the word 'bank', 'The bank is gone' could be true when the building operated by one financial institution is taken over by another; in another sense of the word 'bank', however, the same utterance could be false. It seems to me that this is the case with the word 'energy' as well. In the physicist's sense of the word 'energy', 5 is false; in the common, everyday sense of the word 'energy', however, 5 is true. It wouldn't just be rude or pedantic for somebody to correct my utterance of 5 by pointing out that Daniel weighs more than Zoë; it would be *wrong*. The quantity about which physics theorizes simply isn't what the word 'energy' means in my mouth in a standard utterance of 5. One bit of evidence for this is provided by the fact that we do not feel any pressure to revise our judgment that Zoë is more energetic than Daniel in light of the empirical discovery that energy is proportional to mass. We feel no pressure to respond to the pedant who corrects an utterance of 5 with an appeal to special relativity by thinking *oh, I suppose that I was wrong about Daniel and Zoë*. Rather, we are inclined to respond by saying something like 'you know what I mean'.

Of course, one might feel a *you know what I mean* reaction in two different scenarios: firstly, when one has made a literally false claim, but where a nearby true claim is

[8] Dowe (2000, p. 7)

[9] Two words are polysemous when they share spelling and pronunciation, but have distinct—though related—meanings. Polysemy is distinguished from homonymy, in that the meaning of homonyms—like, *e.g.*, 'bank' (a financial institution) and 'bank' (a riverside)—need not have any relation to one another.

[10] 'Word' appears to be polysemous as well. We might use the word 'word' in such a way that there are (at least) two words 'bank', or we could use it in such a way that there is at most one word 'bank'. Here, I'll use the word 'word' in the second way. In this sense, words are roughly individuated by their phonetic and orthographic features in a way that is blind to meaning.

available and it is clear from context that it was the latter, and not the former, which was intended. For instance, if somebody says 'David Cameron is the president of the United Kingdom', and a pedant corrects them by pointing out that David Cameron is actually the *prime minister* of the United Kingdom, they might well respond by saying something like 'you know what I mean.' However, we do not feel as though the pedant was *misinterpreting* them; the original claim was false (or perhaps truth-valueless)—David Cameron is not the president of the U.K. In this case, the *you know what I mean* reaction is rooted in the feeling that the pedant is to some extent not being a cooperative interlocutor, that they are more interested in nit-picking than in advancing the conversation. Secondly, we might have a *you know what I mean* reaction in cases where we feel that our original claim was indeed true and maximally informative when properly understood, but our interlocutors are, through negligence or malice, *mis*understanding our claim. For instance, when a comic corrects my utterance of 'I should get a hair cut' with 'you should probably get *all* of them cut', the *you know what I mean* reaction is rooted in the feeling that what I said was both literally true and maximally informative, but the comic deliberately misunderstood the claim that I was making by confusing it with another which has the same surface structure. I believe that the pedant who corrects an utterance of 5 with an appeal to special relativity elicits this second kind of *you know what I mean* reaction. There simply was no error in the utterance of 5, any more than there was in an utterance of 'I should get a hair cut'. The pedant, like the comic, is confusing the claim made—which is, as is clear from context, a claim about Zoë's and Daniel's respective personalities—with another claim, a scientific claim, which could be expressed by another sentence with the same surface structure. (There will doubtless be *you know what I mean* reactions which don't fall neatly into either of these two camps. I don't mean to suggest that we have perfect introspective access to whether any particular *you know what I mean* reaction falls into the first or the second camp. It is enough, for my purposes, that there are clear cases, and that the case of somebody correcting 5 with an appeal to special relativity falls clearly into the second camp.)

So I don't think that it's right to say that, in the case of 'energy', an imprecise folk term was *replaced* by a precise physical term. The right thing to say is that, in the development of physics, a term polysemous with the folk term 'energy' was developed. Just as we should not say that the claim that racial discrimination is against the law has been refuted by mature physics—which tells us that racial discrimination is entirely consistent with Schrödinger's equation—we should not say that claims about how much energy Zoë and Daniel have are refuted by mature physics. We were led to this conclusion by considering our inclination to revise judgments like 5 in the light of *a posteriori* discoveries about the referent of the physical term 'energy'. We were not inclined in the least to revise the judgment 5 upon discovering that Daniel has more mass than Zoë and that physical energy is proportional to mass; on this basis, we concluded that the meaning of 'energy' in 5 must be distinguished from the meaning of 'energy' in the sentence 'energy is proportional to mass'.

I believe that a similar argument would apply to any purported empirical analysis of causation which diverged radically from our characteristic causal judgments. Suppose, for the sake of argument, that Dowe (2000) persuades us that there is a physical property—roughly, the transmission or retention of a conserved quantity—which is natural and theoretically useful enough to deserve mention, and which bears enough of a family resemblance to our ordinary talk of 'causation' to merit the name. Call this relation $\langle$CAUSATION$^+\rangle$. Suppose that we then discover that $\langle$CAUSATION$^+\rangle$ does not relate the Surgeon General's report and the subsequent drop in rates of lung cancer—*i.e.*, suppose that there's no appropriate transfer of energy-momentum between the Surgeon General's report and the lung cancer rate. Perhaps that's because this is a case of prevention, where cigarette smoking *would* have caused$^+$ the cancer rates to be high, but the Surgeon General's report simply lowered cigarette smoking, thereby preventing it from having this effect$^+$, without itself transmitting any conserved quantity to the cancer rates. (Or perhaps it's because it's not even clear what it *means* to say that there's an exchange of a conserved quantity between the Surgeon General's report and the lung cancer rates.[11]) In this case, I think that we would, and should, nevertheless hold firm to the judgment that 6 is true.

6. The Surgeon General's report caused lung cancer rates to drop.

In fact, this is exactly what Dowe (2000) does; he takes examples like this as a reason to say that $\langle$CAUSATION$\rangle$ (the relation picked out by our ordinary concept CAUSATION) is the relation that two events bear to one another iff *either* there is the retention or transfer of a conserved quantity from the one to the other *or* there is an appropriate counterfactual relationship between the two events.[12] Call the relation that two events bear to one another iff there is the appropriate counterfactual relation-

[11] See Woodward (2003)

[12] For the details, see Dowe (2000, ch. 6).

ship between them '⟨CAUSATION*⟩'.[13] Then, DOWE claims that our ordinary concept of causation picks out a relation which is the union of genuine *physical* causation, ⟨CAUSATION$^+$⟩ (retention or transfer of conserved quantity), and the counterfactual relation ⟨CAUSATION*⟩.

$$\langle \text{CAUSATION} \rangle = \langle \text{CAUSATION}^{+} \rangle \cup \langle \text{CAUSATION}^{*} \rangle$$

It is entirely unclear, however, why DOWE should feel any pressure to introduce the relation ⟨CAUSATION*⟩ at all. Given DOWE's methodology, what reason do we have for thinking that claims like 6 are true in the first place? Why shouldn't we regard claims like 6 the same way that we regard pre-theoretical physical intuitions like 'the meter stick remains one meter long even when you're moving relative to it'? It is intuitive that the Surgeon General's report caused the lung cancer rates to drop, but so too is it intuitive that the meter stick does not become shorter when it is thrown through the air.

The physicist's ENERGY is a theoretical concept. Assuming a standard semantics for theoretical terms/concepts,[14] the concept ENERGY refers, if at all, to the thing which plays (well enough) the role that ENERGY plays in our theory—that is, it refers, if at all, to the thing that is (near enough) proportional to the relativistic mass times the square of the speed of light, the thing that is (near enough) conserved in closed systems, and so on and so forth. That is to say: we use the concept ENERGY with a great deal of deference to the actual world. It is for this reason that empirical investigation can reveal *a posteriori* facts about the physical property ⟨ENERGY⟩—for instance, that Daniel has more ⟨ENERGY⟩ than Zoë. It is for this reason that it would be foolish to attempt to learn anything interesting about the physicist's ⟨ENERGY⟩ through armchair reflection on thought experiments.[15] We would not, and ought not, respond to the fact that it is very intuitive to say that Zoë has more energy than Daniel by emending our physical theory of energy, saying that the quantity picked out by our everyday concept ⟨ENERGY⟩ is actually a function both of ⟨ENERGY$^+$⟩ (the quantity picked out

[13] Here, I am departing slightly from DOWE's terminology. DOWE used the term 'cause*' for the union of the retention or transference of a conserved quantity and the counterfactual relation.

[14] See, *e.g.*, LEWIS (1970, 1972).

[15] Note, however, that even a theoretical concept like ENERGY can be to some degree investigated through thought experiments. For instance, assuming the standard story about theoretical terms, if there were, at the actual world, no property which played any of the theoretical roles of energy, then there would be no energy—ENERGY, like PHLOGISTON and ORIGINAL SIN, would be a non-referring concept. And we can learn *this* fact about ENERGY from the armchair.

by the theoretical concept) and ⟨ENERGY*⟩ (a quantity corresponding to how excitable a person is).

The point here is just that the way that we do, and ought to, treat the concept CAUSATION (and, indeed, the way that DOWE himself treats the concept CAUSATION)[16] is not the way that we do, and ought to, treat theoretical concepts like ENERGY. This is because CAUSATION is not a theoretical concept. Indeed, DOWE explicitly affirms that "'cause' is not a technically defined term in any scientific theory" (p. 9). We are led to the same conclusion by RUSSELL (1912), who persuasively argues that causal notions do not appear in our fundamental physical theories; when we look to fundamental physics, what we find are not laws formulated in causal terms, but rather differential equations governing the time-evolution of systems in an entirely time-asymmetric, and therefore, acausal, manner. (This is not to deny that we may learn causal facts by studying fundamental physical theories, nor that causal considerations might play into which theories we end up accepting,[17] but it is to deny that those theories employ as a theoretical term 'causation' or any of its cognates). NORTON (2007) additionally argues against the suggestion that some notion of causality is built into fundamental physical theories by providing an ingenious example in which a ball at rest at the tip of a dome begins to roll down the side of the dome without any force being applied to it—thus, without any *cause*—but nevertheless in accordance with Newtonian mechanics (given a certain force law). NORTON's conclusion: the differential equations of Newtonian mechanics do not hide within them any causal notions, since they are compatible with radical failures of causality (events having no causes whatsoever).[18]

Of course, simply because CAUSATION is not a theoretical concept, this does not preclude DOWE from defining up a theoretical notion of *physical causation*, or CAUSATION$^+$; however, if that is his game, then he should not say that his goal is to investigate the "the language-independent entity *called* 'causation'",[19] since the language-

[16] DOWE (2000) additionally defers to our ordinary concept of causation in chapter 7, where he takes the fact that it sounds incredibly odd to say things like 'hitting the tennis ball caused the wall to remain in place' as a reason to revise his theory of physical causation (since there is an exchange of a conserved quantity between hitting the tennis ball, which rebounded off the wall, and the wall's remaining in place).

[17] See, *e.g.*, SMITH (2007).

[18] Though see MALAMENT (2008), who shows that the force law NORTON (2007) uses is only once continuously differentiable at the summit of the dome. MALAMENT expresses some hesitation at saying that such a force law is properly 'Newtonian.' I believe that we could make the same point that NORTON (2007) makes with his dome with reference to so-called 'space invaders'—see (EARMAN, 1986, p. 34); though here, too, there is room to quibble.

[19] DOWE (2000, p. 7, emphasis added). Although, as Daniel Drucker pointed out to me, DOWE

independent entity which is *called* causation is the relation which is picked out by our ordinary concept CAUSATION. Moreover, it is entirely unclear what the criteria of success for such an undertaking would be. Theoretical concepts like ENERGY, CHARGE, and so on earn their keep in making predictions. We can learn about the theoretical property of CHARGE by looking at the predictions that the physical theory implicitly defining CHARGE makes, and then checking to see whether they are borne out. However, Dowe (2000)'s empirical analysis of physical causation makes no novel predictions—Dowe is not doing physics; he is doing metaphysics—so it is unclear how we are supposed to get evidence that our empirical analysis is on the right track, if not by comparing its verdicts to our well-informed characteristic causal judgments. For instance, suppose that we define up a theoretical notion, CAUSATION^{+}, which relates two events c and e iff c is in the past or future light cone of e. This is a fairly natural notion, one which marks a fairly joint-carving distinction between events which are space-like and time-like separated (*i.e.*, two events are causally^{+} related just in case they are time-like separated). Does this represent an adequate empirical analysis of causation? Or does the fact that effects^{+} often precede their causes^{+} doom this analysis?

I take it that this is roughly the point Lewis is making when he writes that

> ...without [conceptual analysis], I see no possible way to establish that any feature of the world does or does not deserve a name drawn from our traditional...vocabulary.[20]

Dowe's response to this objection is to insist that "the word 'cause' as scientists use it...must make some 'historical' or 'genealogical' connection to everyday language."[21] However, the notion of CAUSATION^{+} defined above (CAUSE$^{+}(c, e)$ iff c is in the past or future light cone of e) certainly does bear *some* genealogical relationship to the standard notion CAUSATION—it even does some work explaining the distinction some physicists draw between *causal* and *noncausal* processes in special relativity. So too does Dowe (2000)'s notion of physical causation as the retention or transfer of a conserved quantity. So too does Fair (1979)'s notion of physical causation as the flow of energy. Without any more to anchor our account—without any further idea

(2000) might have merely meant that his goal was to investigate the language-independent entity called 'causation' *by him* (and not necessarily any other English speaker).

[20] Lewis (1994b, p. 415)

[21] Dowe (2000, p. 9)

about what is to constitute success and what it is constitute failure—how are we to decide between these empirical analyses of causality? In practice, Dowe criticizes opposing empirical analyses of causation on two grounds: firstly, that they are conceptually confused or circular (*e.g.*, he criticizes Salmon (1984)'s theory by pointing out that 'object' is defined in terms of 'causation', and 'causation' is defined in terms of 'object'); and secondly, that they commit us to claims about 'causation' that we are unwilling to accept. For instance, he criticizes Fair's theory that causation is the transfer of energy or momentum on the grounds that a spaceship moving inertially with a constant velocity does not transfer any energy or momentum. However, Dowe alleges, its earlier velocity is a cause of its later velocity. I have a hard time understanding this as anything other than conceptual analysis. From whence do we secure the putative datum 'the spaceship's earlier velocity caused its later velocity', if not our implicit grasp of the concept causation? As Dowe recognizes,[22] 'cause' is not a theoretical term in any scientific theory, so the verdict that the spaceship's earlier velocity caused its later velocity cannot be something that we learn from any scientific theory alone. For another example, in the middle of his introductory methodological chapter, Dowe claims that quantum mechanics has disproven the dictum that 'every event has a sufficient cause', since we

> ...are forced to accept that there are cases that we cannot but call 'causation', where the full cause is not a sufficient condition for the effect.[23]

It is difficult to see this, too, as anything other than conceptual analysis. Why can't we help but to call these cases 'causation', if not for the fact that our concept causation intuitively and obviously relates them?

In part, this argument is a *tu quoque*—that Dowe himself utilizes the methodology that he criticizes—but additionally, and more importantly, it is an argument against the possibility of engaging in the kind of project that Dowe sets for himself without leaning on at least *some* judgments about which events are causally related. Without these judgments to constrain our theorizing, there are just too many notions of physical causation to be defined, and no reason to favor one over the others. Moreover, once we open the door to some of our characteristic causal judgments, we would need some principled reason to keep the others out. Else, we might be fairly accused of accepting the characteristic causal judgments which speak against opposing theories

[22] "...'cause' is not a technically defined term in any scientific theory" Dowe (2000, p. 9)
[23] Dowe (2000, p. 8, fn. 9)

and rejecting the characteristic causal judgments which speak against our own theory on an *ad hoc* basis. A proponent of an empirical analysis of causation might want to say that their account appeals to the characteristic causal judgments of scientists (or, perhaps, physicists).[24] However, physicists are also competent speakers of English, so they should have all the same characteristic causal judgments as the rest of us. In order to cordon off some of their characteristic causal judgments, we will have to say something about which of their causal judgments are appropriately *scientific* or *physical*. Not only would we have to do this, but we would have to explain why the judgments of physicists should hold special sway, especially since we've admitted that CAUSATION is not a theoretical concept that one learns by learning any scientific theory. Physicists do, of course, have the advantage of being *better informed* than many of us about the underlying fundamental physical state of the world and the underlying fundamental physical laws of nature. But nobody should have been claiming—and LEWIS certainly wasn't claiming—that we should build a theory of causation around the characteristic causal judgments of the scientifically illiterate. The appropriate data of a theory of causation are the characteristic causal judgments of those who are informed of all relevant details of the case, including (if they are relevant) the fundamental physical state of the system involved, and the fundamental physical laws of nature.

So what I think is right about DOWE's methodology—what remains once the rhetoric about distinguishing the concept of causation from what causation is in the world has been shed—is that a theory of causation should not be blind to what our best theories tell us about the world we occupy. In part, that's because facts about what causes what at the actual world depend upon the nature of the fundamental physical state of the actual world, and the fundamental physical laws of nature at the actual world.

A brief qualification: there may be an extra project to be undertaken, after a philosophical account of causation has been provided; namely, the project of studying whether there is some physical relation obtaining between any two events which are causally related at the actual world. This will be a largely empirical project, since empirical investigation is needed to know which events at the actual world are causally related, and also which physical relations obtain between those events. Now, there may be some quantity—we can call it *biff*, if we like—which is given by causes to their effects. Then again, there may not be. Whether there is biff is in part a question

[24] *Some* proponent of an empirical analysis of causation might want to say this, but DOWE explicitly denies that this is his project. See p. 10–11.

to be decided by scientific investigation—we need to know which events give and take fundamental physical quantities to and from which other events—and in part a question to be decided by philosophical analysis—we need to know, given a full description of the state of the world, which pairs of events are causally related, and to do this, we need a philosophical analysis of causation. If it turns out that there is some quantity that all and only causes give to all and only their effects, then there is biff. If not, then not. In fact, I think that our world is not a biffy-world. Ours is a world containing double prevention—where, *e.g.*, c prevents d from preventing e, and thereby causes e, without transferring any physical quantity to e—so there is no physical property that all causes give to their effects. There may be some more complicated relationship between the causal relation and some kind of biff—see, *e.g.*, Handfield et al. (2008)—but, in any case, the way to investigate the relationship is to first discover which events are causally related, and then to see whether the causal relation corresponds to any fundamental physical relation at the actual world.

1.3 Conceptual Analysis and Truth

Suppose that we find the arguments of the previous section compelling. We might still find it mysterious that armchair philosophical reflection could teach us anything about the *relation* of causation. After all, aren't facts about the relation of causation facts about the physical world? And don't we have to get up out of the armchair in order to discover facts like this? In this section, I want to briefly sketch an epistemology for the philosophical investigation of causation. In brief, my proposal is that for concepts like CAUSATION, the fact that this concept picks out the causal relation, ⟨CAUSATION⟩, is explained by the features of the concept, rather than features of the actual world. For this reason, learning things about the concept can tell us which relation it picks out. Such an epistemology is far from mysterious, I contend, because learning things about the concept CAUSATION doesn't reveal anything more about the relation ⟨CAUSATION⟩ than that it is the relation picked out by our concept—and this is precisely the kind of knowledge that one would expect to come along with conceptual competence.

1.3.1 Concepts and Conceptions

In this section, I will say a bit about what I take concepts to be and argue that we must distinguish concepts from what I will call *conceptions*. This distinction will prove methodologically important, since, even though the *concept* CAUSATION applies to a pair of events, this does not mean that a particular *conception* of causation applies to a pair of events. So, while facts about the *concept* of causation provide data that a philosophical account of causation should feel theoretical pressure to capture, this does not mean that even widespread agreement in people's *conception* of causation ought to constrain our theorizing.

A concept is some kind of mental representation which is in some sense normatively governed by its application conditions. The concept is applied to some things correctly, and applied to others incorrectly. When I categorize Cujo as a dog—when I apply the concept DOG to Cujo—I do so correctly. When I categorize Cujo as a rat—when I apply the concept RAT to Cujo—I do so incorrectly. For some concepts, possession of those concepts guarantees a covert understanding of at least some of the concept's application conditions.[25] (I'll distinguish between beliefs which are *overt* and those which are *covert*. Overt beliefs are those, roughly, which a person is capable of explicitly formulating and sincerely affirming; whereas covert beliefs are, roughly, those which they are not capable of either explicitly formulating or sincerely affirming. For instance, speakers of a natural language covertly belief—and covertly know—the rules of universal grammar. However, the vast majority are incapable of formulating the rules of universal grammar, and may even explicitly deny such rules when they are presented with them.) This does not mean that those who posses a concept always apply it correctly. For instance, I might apply the concept PRESIDENT to a Barack Obama look-alike. That does not mean that I don't possess the concept PRESIDENT—merely that I am mistaken about certain empirical facts about the person before me, or perhaps that I have just woken up, ingested drugs, or am otherwise not in an optimal state of mind. Nevertheless, for a variety of concepts in a variety of circumstances, if a person possesses a concept, then when they are fully informed of the relevant details and thinking clearly, they will be in a position to make correct judgments about whether or not the concept applies.[26] It is for this reason that philosophical thought

[25] In terms of the epistemic two-dimensional framework which I will discuss below, I believe that possessing a concept puts us in a position to have a covert understanding of the concept's *two-dimensional intension*.

[26] I don't take these claims to be particularly controversial. They are, for instance, consistent with the so-called classical theory of concepts, according to which concepts have implicit definitions—necessary and sufficient conditions for their application. They are also consistent with the prototype theory of concepts, as well as conceptual atomism. For a discussion of philosophical and psychological theories of concepts, see MARGOLIS & LAURENCE (1999).

experiments provide a useful arena for exploring the application conditions of some concepts. Take, for instance, the concept KNOW, expressed by the English verb *know*. When we consider a situation in which, at 12:00, John looks at a broken clock reading '12:00' and forms the belief that the time is 12:00, we judge that John does not know that the time is 12:00. That is, we judge that the English sentence 'John knows that the time is 12:00' is false and that its negation is true. From this judgment we can infer that the concept KNOW does not apply to John and the proposition ⟨the time is 12:00⟩ in this situation.[27] Call judgments like these—immediate, unreflective judgments about the appropriateness of English sentences involving the words which express the concepts in question—*characteristic judgments* about the concept's application conditions.

It is my view that our *characteristic causal judgments* can constitute evidence for or against a theory of causation. If we have a clear causal judgment that c caused e, that is a *prima facie* mark against a theory which says that c didn't cause e. Of course, there may be other theoretical virtues which lead us to accept a theory of causation which doesn't align with all of our characteristic causal judgments. Nevertheless, all else being equal, a theory which accounts for more of our characteristic causal judgments is *ipso facto* a better theory.

Five qualifications are in order. Firstly, I do not suppose that our characteristic judgments are necessarily judgments about the *truth* or *falsity* of certain English sentences. Rather, I am merely supposing that they are judgments about the *appropriateness* of English sentences. It is always possible for us to claim that a characteristic judgment that an English sentence is inappropriate is due not to any *semantic* defect of the sentence, but rather to some *pragmatic* defect. For instance, the sentence

7. Every libertarian president of the U.S. has paid down the national debt.

will be judged inappropriate by most competent and politically informed English speakers. However, most popular theories of quantification will count 7 as vacuously true; since there have been no libertarian presidents of the U.S., all of the none of them have paid down the national debt. If we accept these theories of quantification, then we will explain the inappropriateness of 7 by appeal to pragmatic theories—for

[27] I'm assuming that ⟨the time is 12:00⟩ is a proposition, rather than, *e.g.*, a property attribution (*c.f.* LEWIS, 1979a). Nothing hangs on this assumption. If you think otherwise, have John give a name to the present moment in our example—say he calls it 'Jimbo'—and then let the proposition that John fails to know be the one expressed by the sentence 'Jimbo is 12:00'.

instance, by Grice's maxim of quality, the claim 'All Fs are G' implicates that there exist some Fs.

Secondly, there may very well be one word in our language which expresses two distinct but closely related concepts—*e.g.*, as is the case with 'bank', the financial institution and 'bank', the building in which such a financial institution operates. I do not suppose that we have overt introspective access to whether a single word in our language is polysemous in this way. Philosophical and/or linguistic reflection may be required in order to bring out this kind of polysemy. Indeed, this is almost certainly the case with 'cause', as this word is sometimes used to mean 'because', as in 'He went to the store, cause he was out of milk.' Of course, not much reflection is required to recognize this polysemy (the phonetics of the two words are even different in American English; the explanatory 'cause' is pronounced /kuhz/, whereas the causal 'cause' is pronounced /kawz/). More interestingly, Hall (2004) argues that there are two concepts of causation—which he dubs 'production' and 'dependence'—both of which are expressed by the word 'cause'. Though I disagree with Hall (2004) about this point, I don't mean to rule his position out as a matter of methodology.

Thirdly, I do not suppose that *every* concept is such that possessing that concept puts us in a position to have a covert understanding of that concept's application conditions. It could be that some concepts cannot be individuated by their application conditions, but rather only by the role they play in our cognitive lives. For instance, some metaethical expressivists take the concept OUGHT to function merely as a way of expressing non-representational mental states like *being for* some action. For such expressivists, a person can possess the concept OUGHT without having knowledge of OUGHT's application conditions, simply because they haven't yet made up their mind about what they are for.[28] I am not disagreeing with those metaethical expressivists (at least, not here). I am merely claiming that CAUSATION is not like OUGHT in this respect. To possess the concept CAUSATION is to be in a position to be able to correctly judge which causal claims are appropriate in a given scenario.

Fourthly: I say that to possess the concept CAUSATION is to be in a position to be able to correctly judge which causal claims are appropriate in a given scenario. I do not say, however, that possessing the concept CAUSATION is sufficient for being able to correctly judge which causal claims are appropriate in a given scenario. In general, I do not suppose that concept possession alone is sufficient for correctly categorizing

[28] I take it that this isn't the view of more sophisticated *Quasi-Realists* like Allan Gibbard (2003).

the things to which the concept applies. As WILLIAMSON (2007) stresses, it is usually the case that abilities over and above mere concept possession are necessary for the correct application of the concept. Figuring out whether a given figure is a chilliagon, for instance, requires either a good enough memory to count up the sides of the figure, or else the ability to keep track of the number of sides in some other fashion. These abilities are not guaranteed by possessing the concept CILLIAGON. The googol-digit number which appears in the first googol digits in the decimal expansion of π may or may not be prime. We are not able to categorize it as prime or composite, but this doesn't mean that we don't possess the concept PRIME.

Finally, and most importantly, our characteristic judgments about the application of a concept should not be confused with any overt beliefs we may have formed about the referent of that concept. Call our overt beliefs about the referent of a concept our *conception* of that referent. A conception is something like a theory about the referent of our concept—though, in the case of most people, such conceptions will be at best proto-theories. Conceptions often come into conflict with characteristic judgments. For instance, some philosophers once had a conception of knowledge according to which S knows that p iff S has a justified, true belief that p. The case of John and the broken clock given above, however, shows that our characteristic judgments about KNOW fail to line up with this conception. Our concept KNOW does not apply to John and the proposition ⟨the time is 12:00⟩, even though John has a justified, true belief that the time is 12:00. Philosophers are not the only ones with conceptions. Anyone who has taken or taught an undergraduate Epistemology course knows that many undergraduates have, or can be easily led to adopt, a conception of knowledge according to which you can only know that p if your subjective experience is incompatible with every possibility in which $\neg p$. They can be rather easily led, via this conception, to embrace the conclusion that nobody knows anything about the external world. Nevertheless, these students continue to make characteristic judgments according to which I know what will be on the final, how it will be curved, *etc*.

More surprisingly, recent experimental results[29] have shown that, if, for instance, you present subjects with the John case above, and ask them whether John knows that the time is 12:00, a significant portion of them will systematically judge that John *does* know that the time is 12:00. If we take people's responses to surveys to reliably tell us something about their characteristic judgments, then we seem to be

[29] See, *e.g.*, STARMANS & FRIEDMAN (2012).

led to the conclusion that the philosophical account according to which knowledge is justified true belief is not in tension with people's characteristic causal judgments, contrary to Russell (1948), Gettier (1963), and the widespread consensus of the philosophical community. I do not think that we ought to respond to such evidence by concluding that the concept KNOW does apply to John and the proposition ⟨the time is 12:00⟩. The respondents in such a survey are surely making a mistake—albeit an apparently widespread and predictable mistake. If there were any doubt about this before, these doubts ought to have been laid to rest by Turri (2013), who shows that, by presenting cases like these more slowly and carefully, you can get respondents to correctly judge that John doesn't know that the time is 12:00. So respondents are, in the original studies, making a mistake. What's needed is an explanation of why they systemically and predictably make this mistake.

Williamson (2007) explains mistakes like these by appealing to my fourth qualification above—namely, that, for concepts like KNOW, concept possession does not on its own guarantee the ability to correctly apply the concept. Thus, Williamson's explanation of the odd judgments about John and the broken clock is that, though respondents possess the concept KNOWS, they do not yet have the abilities necessary to correctly apply that concept to cases like John and the broken clock. The situation with these respondents and the judgment that John knows that the time is 12:00 is analogous to the situation with non-linguists and the judgment that 'The rat the cat the dog scared chased escaped' is not grammatical.[30] Though non-linguists have covert knowledge of the rules of UG, they do not always have the ability to apply those rules, especially in eccentric and complicated sentences involving many nested noun phrases. Similarly, though respondents have covert knowledge of the application conditions of KNOW, they do not always have the ability to correctly classify John's situation as a case of non-knowledge.

I agree that, in general, abilities over and above concept possession are necessary for correct application of the concept. However, Williamson's explanation of the mistake involved in the case of John and the broken clock sells non-philosophers a bit short. Figuring out that you can't learn the time from a broken clock is nowhere near as cognitively demanding as figuring out that 'The rat the cat the dog scared chased escaped' is grammatical, or determining the number of sides on a chilliagon. It does not take a philosophy course to teach most people that you can't learn the time from

[30] The sentence means that the rat which was chased by the cat which was scared by the dog escaped.

a broken clock.

In my judgment, the fifth and final qualification above does a much better job diagnosing the error which respondents make when they falsely claim that John knows the time. The distinction between concepts and conceptions affords us the hypothesis that, when respondents encounter the case of John and the broken clock, they are clued in to the fact that there is something peculiar or idiosyncratic about the case. The prompts generally stress the fact that John's belief is *true*, for instance (presumably to make sure that respondents recognize this fact before responding). Respondents are thus led to wonder whether, given that such features are being stressed, they might be important in some way. They stop providing their unreflective, covert judgments, and start overtly thinking through the relevance of the truth of John's belief to the question of whether John knows that the time is 12:00. And the only way to overtly think through this question is by appeal to one's overt beliefs about what it takes for somebody to know that some proposition is true—that is to say, by appeal to some *conception* of knowledge. They may, for instance, reason along the following lines: *John did use the clock to arrive at a* true *belief, and isn't that just what all of us do all the time when we look at working clocks? And, after all, what's the difference between the state of the clock when it's working and when it's not? Surely it's just whether the clock tells the* correct *time; and in this case, the clock* does *tell the correct time. So, the clock is just like a working clock in all relevant respects when John is looking at it; so if any of us ever know the time by looking at a working clock, then John must know that the time is 12:00.* Overt reasoning like this makes use at several points of overt beliefs about the relation ⟨KNOW⟩. Such beliefs are part of the respondent's *conception* of knowledge. So, if the respondent is clued into the fact that something is awry with the case, and they therefore stop responding unreflectively, but rather appeal to some kind of overt reasoning like the foregoing, then we ought to expect their responses to only tell us something about their conception, and not their concept, of knowledge. But, as we saw above, *conceptions* and *concepts* can easily come apart. So there's no reason to take what such respondents say as evidence that the agents in Gettier cases have knowledge.

This is, of course, just a hypothesis about what's going on with these survey results; it could be tested by measuring how much time respondents spend on this question, compared with how much time they spend on more mundane questions. Alternatively, we might test the hypothesis as follows: present subjects with a vignette in which they are told that the clock is broken, reads '12:00', and John, falsely believing that it is working, looks at the broken clock and forms the belief that the time is

12:00. Then ask them whether a) John knows what time it is, b) John doesn't know what time it is, or c) there isn't enough information to say. The hypothesis that people's concept of knowledge is such that people know in the Gettier case would predict (c). They have been told that John has a justified belief that the time is 12:00. All that is needed in order to transform this into a Gettier case is truth. So, if agents have knowledge in the Gettier case, then the respondents ought to answer (c). My hypothesis, on the other hand, predicts that people would answer (b). They have an implicit grasp of the concept KNOW and are therefore able to see that one cannot come to know the time by looking at a broken clock. Since the vignettes do not place any particular emphasis on unusual features of the case, respondents shouldn't have any reason to engage in a process of overt reasoning about knowledge by appealing to their conception of knowledge. So, they will covertly recognize that the fact that the clock is broken is enough to establish that John doesn't know the time.

1.3.2 Concepts and Metasemantics

Consider two concepts: ROUND and WATER. For both of these concepts, we should accept the trivial claims (8) and (9).

8. ROUND refers to ⟨ROUND⟩.

9. WATER refers to ⟨WATER⟩.

The concept ROUND refers to the property of being round; ROUND applies to an entity just in case that entity is round. The concept WATER refers to the property of being water; it applies to a material just in case that material is water—we now know that this will hold iff that material is made up of molecules consisting of two parts hydrogen and one part oxygen.

On the assumption that these concepts refer at all, these claims are trivial, on a par with the claim that $\langle p \rangle$ is true iff p, which is trivial on the assumption that $\ulcorner p \urcorner$ is meaningful. (Part of my goal in this section is to explain their triviality.) Nevertheless, there is an interesting difference between (8) and (9). Roughly, the difference is this: the fact that (8) is true is adequately explained by facts about the concept ROUND and the property ⟨ROUND⟩ alone—*without reference to any other facts*. In order to know that ROUND picks out the property ⟨ROUND⟩, we don't need to know anything, for instance, about which world is actual. That's because we don't use the concept ROUND with deference to the actual world. People in other possible worlds could use our concept ROUND to pick out the very same property ⟨ROUND⟩. Nor does it matter whether there even *are* any round things here at the actual world—this would not render (8) false. WATER is different. We *do* use the concept WATER with deference to the actual world. The fact that (9) is true is not adequately explained by facts about the concept WATER and facts about the property ⟨WATER⟩. In order to properly explain (9), we must additionally point to facts about the actual world—namely, the fact that the watery stuff filling the lakes and rivers in the actual world is composed of molecules containing two hydrogen atoms and one oxygen atom. If the actual world had turned out to be different—if, for instance, the watery stuff filling the lakes and rivers in the actual world had turned out to be composed of molecules containing two hydrogen atoms and two oxygen atoms, then WATER would have referred to ⟨HYDROGEN PEROXIDE⟩. People in possible worlds like these would use our concept WATER to refer to ⟨HYDROGEN PEROXIDE⟩.[31]

We might want to draw this distinction with reference to the framework of epistemic two-dimensional modal semantics.[32] The rough outline of this framework is as follows. We have a set of *epistemic possibilities*—(centered) ways the actual world might turn out to be, for all we know *a priori*. We endorse a kind of *scrutability thesis* to the effect that, given a world in this set—*i.e.*, given a specification of which world is actual—we have a (perhaps covert) understanding of what the meaning of a concept or a term in our language would be, were that possibility actual. I'll assume that the meaning of a concept or a term in our language is provided by its *intension*—that is, by a function from worlds to extensions, or sets of (n-tuples of) objects to which the concept or term applies at that world.

To make this a bit more precise, we may introduce a doubly-indexed interpretation function $[\![\ ]\!]^{\circ,*}$. The first index, $\circ$, is filled by an epistemic possibility—a (centered) way the actual world might be, for all we know *a priori*. The second index, $*$, is filled by a possible world. The *two-dimensional intension* of a concept C will then be a

[31] I'm here supposing, for the sake of illustration, that WATER is individuated roughly by the application condition 'WATER applies to an entity if and only if that entity belongs to the same natural kind as the watery stuff filling the lakes and rivers around here in the world that I occupy.' That is: I'm assuming that the concept WATER has a kind of rigidified descriptive content. I think that this is roughly correct, and I'm using this assumption to help illustrate the distinction between concepts like ROUND and WATER, but I don't believe that I need this assumption for anything other than illustrative purposes. What I *do* need is the assumption that, given a certain description of a way the actual world might be, we are able to understand how our concept applies in that scenario.

[32] See CHALMERS (2006a,b).

function $[\![\textsc{c}]\!]^{\circ,*}$ from epistemic possibilities to functions from worlds to extensions. The *intension* of a concept $\textsc{c}$ will be a function from worlds to extensions. The function $[\![\textsc{c}]\!]^{\circ,*}$ gives us the two-dimensional intension of $\textsc{c}$; whereas, for any i, the function $[\![\textsc{c}]\!]^{i,*}$ gives us the intension of $\textsc{c}$, if i is actual.

In this framework, the interesting distinction between ROUND and WATER is that the two-dimensional intension of ROUND is constant, that is,

$$\forall i, j \; [\![\textsc{round}]\!]^{i,*} = [\![\textsc{round}]\!]^{j,*}$$

whereas the two-dimensional intension of WATER is not constant.

$$\exists i, j \; [\![\textsc{water}]\!]^{i,*} \neq [\![\textsc{water}]\!]^{j,*}$$

It is my contention here that, in this respect, CAUSATION is more like ROUND than like WATER. Given full knowledge of a possible world, in order to answer the question of which events caused which other events at that world, I don't have to know which world is actual.[33]

$$\forall i, j \; [\![\textsc{causation}]\!]^{i,*} = [\![\textsc{causation}]\!]^{j,*}$$

Moreover—and this is, I think, the key move—the *relation* of causation, $\langle$ CAUSATION $\rangle$, can be identified with (or at least agrees with) the *intension* of CAUSATION ('agrees with' in the sense that, if c and e are causally related at a world ω, then $< \omega, c, e >$ is within the intension of CAUSATION).

$$[\![\textsc{causation}]\!]^{@,*} = \langle \textsc{causation} \rangle$$

From this it follows that we may know *a priori* facts about the relation of causation, since (by the *scrutability* thesis),[34] we have a covert understanding of the intension of CAUSATION, given that any particular world is actual; and, no matter which world is actual, our concept of causation will pick out the very same intension, which is just the causal relation.

$$\forall i \; [\![\textsc{causation}]\!]^{i,*} = \langle \textsc{causation} \rangle.$$

33 I don't have much by way of argument here other than intuition pumping; think about a causal relation inferred from a work of fiction, *e.g.*, 'Romeo's drinking the poison caused him to die.' Does it really seem that this claim could be undermined by discoveries in physics, the same way that the claim 'Ophelia drowned in H_2O' could be undermined? I invite you to agree that it could not; but I have nothing to say to those who decline the invitation.

34 This thesis is defended at length in Chalmers (2012).

So, it seems to me, if it is granted that CAUSATION is not used with deference—if, in order to know which events caused which other events in a certain possibility, I don't have to know which possibility is actual—then, on the assumption that the relation of causation is (or agrees with) the intension of CAUSATION, it follows that we may have *a priori* knowledge of the causal relation.

I suspect that the resistance to this argument will come with my assumption that the *relation* of causation is, or agrees with, the *intension* of the concept CAUSATION. Why, it will be incredulously asked, would the *relation* ⟨CAUSATION⟩ out in the world line up with the intension of the *concept* CAUSATION? The answer to this question is, I believe, that what we *mean* by the expression 'the relation of causation' is just the thing that agrees with the intension $[\![\text{CAUSATION}]\!]^{@,*}$.

By way of explanation: we might be interested in the question 'why is (8) true?' That is, 'why does ROUND refer to ⟨ROUND⟩?' More specifically, we might be interested in the contrastive question, 'Why does ROUND refer to ⟨ROUND⟩, as opposed to ⟨SQUARE⟩ or ⟨TRIANGULAR⟩ or even ⟨SMOUND⟩ or ⟨BLOUND⟩ or...?' (⟨SMOUND⟩ and ⟨BLOUND⟩ are just gerrymandered properties like 'round in worlds with English speakers and triangular in worlds with French speakers and no English speakers and at least twenty feet from anything red in worlds with no English or French speakers'.) Contention: we may answer this question by only appealing to facts about the concept ROUND itself. On my view, the properties ⟨SQUARE⟩ and ⟨TRIANGULAR⟩ and even ⟨SMOUND⟩ and ⟨BLOUND⟩ and so on and so forth are all out there to be referred to. The reason that ROUND refers to ⟨ROUND⟩ is just that ⟨ROUND⟩ is the property out in the world which answers to the concept ROUND.

Consider, on the other hand, the question of why (9) is true. Consider, that is, the question of why WATER refers to ⟨WATER⟩, as opposed to ⟨HYDROGEN PEROXIDE⟩ or ⟨TWATER⟩ or ⟨SCHWATER⟩ or.... This question *cannot* be answered by appealing to facts about the concept WATER alone. We must also make reference to features of the actual world in order to explain the truth of (9).

Imagine that we have two archers shooting arrows at a wall. For the first archer, a target is painted on the wall, and she then draws back her arrow and shoots. The second draws back his arrow, shoots, and then, wherever his arrow lands, a target is painted. Consider the question 'how could the archer hit the target?' The answer we give to this question will depend upon which archer we are talking about. If we are talking about the first, then we will have to talk about the archer's sensitivity to certain features of the target. If, however, we are talking about the second, then we will not think that any sensitivity to the position of the target was required. *Wherever* his arrow landed, he would have hit his target—for the target was determined by where his arrow landed.

When we deploy the concept WATER, we are like the first archer. Because WATER is used with deference, the actual world has set a certain target for that concept which we are not in a position to fully understand until we have conducted empirical research. However, when we deploy the concept ROUND, we are like the second archer. By acquiring the non-deferential concept ROUND, we thereby implicitly determine the property that it refers to, irrespective of what the actual world is like—that property is just the property which agrees with the intension of ROUND. If CAUSATION were a theoretical concept, then it would be like WATER, and we would not be in a position to have an *a priori* covert understanding of its application conditions merely in virtue of possessing the concept CAUSATION. However, given that CAUSATION is not a theoretical concept, and given that it is not used with deference in any other way, that concept is like ROUND, and not like WATER, in that concept possession puts us in a position to have an *a priori* covert understanding of its application conditions (modulo the five qualifications from §1.3.1 above).

I suspect that much of the skepticism about conceptual analysis stems from the sense that it would be mysterious if reflection on the nature of our concepts was able to reveal something about a relation out in the world. However, on my understanding, this is no more mysterious than the second archer's ability to hit his target. Had we another concept, SCHMAUSATION, *this* concept's intension would have agreed with some *other* relation, ⟨SCHMAUSATION⟩, and *it* would have been the relation that we were talking about when we spoke of 'the relation of schmausation.' We could then have used conceptual analysis to learn about the relation of schmausation. Both of these relations are out there to be talked and thought about; the reason that possessing the concept CAUSATION allows us to think and talk about the first is just that this concept implicitly determines which of the two we are thinking and talking about when we utilize it. Conceptual analysis therefore doesn't reveal anything about the properties or relations our concepts pick out other than that *they are* the properties and relations which our concepts pick out. And this is just the kind of covert knowledge that one would expect to come along with conceptual competence.

A few qualifications: the analogy with the archer is somewhat strained. Firstly, in the case of the concept CAUSATION, there is not just one target, which is painted after we develop the concept; rather, the wall is already painted with many targets, and the one that the concept refers to is determined by where the arrow lands. Secondly, our covert knowledge of the application conditions of the concept may not, by itself, completely settle the question of which relation that concept refers to, because of vagueness or indeterminacy in the concept's application conditions, *e.g.*, the concept may be a kind of *prototype* which simply doesn't settle how the concept is to be applied in certain cases. Indeed, I think that this happens in cases of symmetric overdetermination, see §5.3.9. If that is the case, then considerations of simplicity or other theoretical virtues may be required in order to determine which relation that concept picks out. I believe that such theoretical virtues are *a priori*; however, if they are *a posteriori*, then that part of the project of theorizing about causation would have to be *a posteriori*, too. Finally, nothing I have said here should lead us to think that CAUSATION is a relation *worth thinking and talking about*. It may very well be that the concept CAUSATION bears no interesting relation to the other things that we care about—*e.g.*, explanation, moral responsibility, rational action, and so on. In fact, I believe that CAUSATION is a concept worth thinking and talking about, that it does bear important relations to explanation, moral responsibility, and so on. I believe that this concept is a relatively fundamental part of our psychology, and that its importance and usefulness may be partly explained by evolutionary considerations—creatures with the concept of CAUSATION were better fit to understand and control their world, and therefore more likely to survive and pass on their genes—but I won't be defending that claim here. Here, my goal is merely to allay skeptical doubts about using armchair reflection to learn things about a mind-independent relation.

1.4 In Summation

I have argued, firstly, that to the extent that there is a worthwhile project of empirical analysis to be pursued, it is not the kind of project that may be pursued independently of the project of conceptual analysis. Secondly, I have attempted to allay skeptical doubts about the project of conceptual analysis. On my understanding, there is nothing mysterious about using our covert understanding of certain concepts to come to an understanding of language- and mind-independent properties and relations out in the world, since (so long as those concepts aren't used with deference) *which* properties and relations these concepts refer to is implicitly determined by the concepts themselves. In other words, conceptual analysis doesn't reveal anything about the properties or relations our concepts pick out other than that *they are* the properties and relations which our concepts pick out. And this is exactly the kind of knowledge that one would expect to come along with conceptual competence. Nor is this methodology seriously challenged by recent results in so-called experimental philosophy. We can accommodate these results by distinguishing between *concepts* and *conceptions*. (This is not to deny, by the way, that empirical results can tell us something about our concepts; it is merely to deny that certain empirical results warrant skepticism about the methodology of conceptual analysis). In the subsequent chapters (in particular, chapters 2 and 5), I will be utilizing this methodology to theorize about the conditions under which two events are causally related.

CHAPTER 2

THE REDUCTION, ELIMINATION, AND EMERGENCE OF CAUSATION

2.1 Introduction

In this chapter, I want to foreground a division between philosophical theories of singular causation which has received relatively little attention in the literature. In one camp are those theories according to which the causal relations of our day-to-day, macroscopic lives, as well as the causal relations investigated by special sciences like climatology, economics, and epidemiology, ultimately reduce to causal relations between microphysical events—that the causal relation between Zimbabwe's monetary policy and its hyperinflation is in principle reducible to causal influence between the fundamental particles realizing those events. Following Menzies (1988), call the members of this camp *causal reductionists*. Causal reductionists are opposed on two sides. On one side are those who deny that there are any high-level causal relations to be reduced—those who think that the causal relations between fundamental physical states of the world are all the causal relations that there are. Call the members of this camp *causal eliminativists*. On the other side are those who, like causal reductionists, accept the existence of high-level causal relations between smoking and cancer, carbon emissions and climate change, and monetary policy and inflation; however, unlike the causal reductionists, they deny that these high-level causal relations are reducible to low-level causal relations between fundamental physical events. Call the members of this camp *causal emergentists*.

Strevens (2008) has the following to say about causal emergentism:

> Some philosophers suspect that...there are irreducible high-level causal relations...
> Given what we now know, these suspicions are, I believe, extravagant: there simply are no causal relations of which we are aware that cannot be attributed to to lower-level interaction and, ultimately, to the causal influence of fundamental particle on fundamental particle.[1]

The primary goal of this paper is to persuade you that Strevens is incorrect—that, given what we now know, anything *other* than causal emergentism is extravagant.

In §2.2, I will introduce and clarify the division between theories of causality which are causally reductionist, causally eliminativist, and causally emergentist. I will illustrate the division between causal reductionism and causal emergentism with David Lewis's two accounts of causation, one of which entails causal emergentism, the other of which entails causal reductionism. I will illustrate causal eliminativism with the views of Michael Strevens. In §2.3, I will argue that causal reductionism and causal eliminativism are incapable of adequately accounting for the apparent abundance of causal relations between fine-grained events and the apparent scarcity of causal relations between coarse-grained events. In response, I will counsel a rejection of causal reductionism and causal eliminativism, and an acceptance of causal emergentism. As I will explain in more depth below, causal emergentism is consistent with the thesis of *microphysical reductionism*—the thesis that all facts reduce to microphysical facts. In order to be causal emergentists, we need only deny that high-level causal facts reduce to certain microphysical *causal* facts, and not that they reduce to *any* microphysical facts.

[1] Strevens (2008, p. 82). When Strevens says that high-level causal relations are "reducible", he means that they are reducible to low-level causal relations.

2.2 The Division

In his 1973, David Lewis outlined a counterfactual theory of singular causation. According to this theory, an event e causally depends upon a distinct[2] event c iff, had c not occurred, e would not have occurred either,

$$e \textit{ causally depends upon } c \iff \neg O(c) \mathbin{\Box\!\!\rightarrow} \neg O(e)$$

This counterfactual is evaluated according to the semantics given in Lewis (1973b, 1979b): $\neg O(c) \mathbin{\Box\!\!\rightarrow} \neg O(e)$ is true iff the closest world in which c fails to occur is a world in which e fails to occur as well.[3] Causation is the ancestral, or the transitive closure, of causal dependence. That is, one event c causes another event e iff there is a chain of events running from c to e such that, for each link in the chain, the second event causally depends upon the first.

Because the 1973 account determines whether e causally depends upon c by looking to the closest world at which c fails to occur, it requires a sharp line to be drawn between those possibilities in which an event occurs in a different *manner* and those possibilities in which the event simply fails to occur at all. For instance, suppose that in the actual world, Suzy throws the rock overhand. In order to evaluate whether this throw caused the bottle to shatter, Lewis's 1973 account is forced to legislate on whether a world in which Suzy throws the rock underhand is a world in which Suzy's actual throw occurs, but in a different manner, or whether it is a world in which Suzy's actual throw is replaced by a different throw. In his 2000 revision of the counterfactual account, Lewis expresses doubt about the prospects of drawing this line in general. He thus abandons the strategy of looking to the closest world at which the cause fails to occur, and instead focuses on *all* the possible alterations of the time, manner, or occurrence of an event, whether those alterations constitute the same event or not. Whether Susie's throwing underhand is a different event from her actual throw or not, it will still constitute an *alteration* of her actual throw. Lewis then says that an event c *causally influences* a distinct event e iff there is a substantial range of not-too-distant alterations of c, $c_1 ... c_N$, and a range of alterations of e, $e_1 ... e_N$ (at least some of which differ), such that the alterations of c counterfactually pattern with the alterations of e.

[2] Here, 'distinct' means a bit more than 'non-identical'. It must also be the case that e does not *imply*, and is not implied by, c and that c and e not overlap. See Lewis (1986a).

[3] If we drop the assumption that there is such a world, then $\neg O(c) \mathbin{\Box\!\!\rightarrow} \neg O(e)$ is true just in case there is a possible world in which neither e nor c occurs which is closer than any world in which e occurs and c does not. The standards of similarity are those specified in Lewis (1979b).

That is, c causally influences e iff, for all i, had c_i occurred, e_i would have occurred.

$$c \textit{ causally influences } e \iff O(c_1) \mathrel{\Box\!\!\rightarrow} O(e_1) \wedge \ldots \wedge O(c_N) \mathrel{\Box\!\!\rightarrow} O(e_N)$$

Causation is the ancestral of causal influence. That is, c causes e iff there is a chain of events running from c to e such that, for each link in the chain, the first event causally influences the second.

On the surface, these two accounts appear very similar. In Lewis's terminology, we've simply traded whether-upon-whether counterfactual dependence for whether-, when-, and how-upon-whether, -when, and -how counterfactual dependence. Despite their similarities, there is one dimension along which the two theories differ notably. Lewis's 2000 account entails *causal reductionism*. His 1973 account, on the other hand, entails *causal emergentism*. Allow me to explain.

2.2.1 Causal Reductionism

Some prefatory remarks: in what follows, I will call the entity denoted by a nominalization flanking the verb 'cause' in a causal claim like 'Chris's smoking caused his contraction of cancer' an *event*. I use this word simply because most theories of causation claim that it is events, as opposed to facts, which are causally related. However, I mean to leave it open what the causal relata are. I'll call an event picked out by a nominalization describing a region of spacetime in maximally specific detail with just the predicates of fundamental physics a *finely-individuated*, or *low-level* event. Low-level events pin down precisely what is happening, at a fundamental physical level, within a given region of space-time. Events picked out by nominalizations involving predicates more coarse-grained than this ('chemical reaction', 'birthday party', or 'economic depression', *e.g.*) I'll call *coarsely-individuated* events. If an event is distinct from every low-level event, then I'll call it a *high-level* event. So, if you think that stagflation is just the event of thus-and-such fundamental entities being arranged thus-and-so with thus-and-such fundamental properties over a particular time period, then, even though stagflation is coarsely-individuated, you don't think that it is a high-level event, as I am using that term. That is: I leave it open whether coarsely-individuated events are high- or low-level events. I will assume throughout, however, that if there are high-level events, then most coarsely-individuated events are high-level. Note also that, while disjunctions of nominalizations picking out low-level events only utilize the predicates of fundamental physics, they do not describe a region of spacetime in maximally specific detail; so, the event picked out by such a disjunction will not count as a low-level event unless it is identical to an event picked out by one of its disjuncts. Indeed, one may think that every high-level event is just a (perhaps infinitary) disjunction of low-level events. Instances of the causal relation between high-level events, I'll call *high-level causal relations*, and instances of the causal relation between low-level events, I'll call *low-level causal relations*.

Jaegwon Kim gives voice to the position I'm calling 'causal reductionism' when he writes that

> macro-causality...must be viewed not as something basic and fundamental but as something that is reducible to, and explainable in terms of, more fundamental causal processes.[4]

On Kim's view, the reduction of macro-causality to micro-causality is to be carried out as follows:

> if the macrocausal relation to be reduced is one from an instance of a property F to an instance of property G, we need to correlate F with some micro-property f, and also G with g, and then we show that f and g are appropriately causally connected.[5]

For Kim, an event is just a property exemplification, so when he talks about macro- and micro-properties, he is talking about the macro- and micro-*events* of those properties being exemplified.[6] The kind of correlation between the macro-properties and the micro-properties that Kim has in mind is the relation of supervenience. To correlate F with f is to show that F supervenes upon f. However, for the purposes of characterizing the position I'm going to call causal reductionism, we can remain neutral on what relation it is that high-level events bear to low-level events. Let's just call that relation, whatever it is, 'reduction'. And let's call its inverse 'realization.' Then, we can characterize Kim's view as follows.

> Of nomic necessity, where C reduces to c and E reduces to e, C's causing E reduces to c's causing e

[4] Kim (1984b, p. 51).
[5] Kim (1984a, p. 261).
[6] See Kim (1976).

Throughout, I'll use the uppercase 'C' and 'E' as variables ranging over high-level events, and I'll use the lowercase 'c' and 'e' as variables ranging over low-level events. (Be on guard: in this thesis, 'reduction' is used in two senses: first, as a placeholder for whatever relation the occurrence of high-level events is supposed to bear to the occurrence of low-level events according to the causal reductionist; and second, for the relation between the high-level causal facts and the low-level causal facts in virtue of which they obtain. These could be one and the same relation, but they need not be.) This thesis leaves it open how high-level events are to reduce to low-level events—it could be supervenience, or grounding, or material constitution, or something else altogether.

If, as a matter of nomic necessity, A reduces to B, then the material conditional $B \Rightarrow A$ will be nomically necessary, so the above thesis entails

> *Causal Reductionism*: Of nomic necessity, when C reduces to c and E reduces to e,
>
> $$c \textit{ caused } e \quad \Rightarrow \quad C \textit{ caused } E$$

For the purposes of this paper, I'm going to take this weaker thesis—along with the assumption that some high-level events are causally related—to define *causal reductionism*.[7],[8],[9]

[7] The reason for this is that I am interested in divisions between philosophical theories of causation, and even though many extant theories of causation entail either *causal reductionism* or its negation, those which entail *causal reductionism* are not usually up front about whether they take C's causing E to *reduce* to c's causing e or merely be nomically necessitated by c's causing e.

[8] You could accept the thesis I've labeled *causal reductionism* while denying the nomological possibility of low-level causation, yet accepting the existence of high-level causation. Then, because you would think that the antecedent of the material conditional 'c caused $e \Rightarrow C$ caused E' is false at every nomologically possible world, you would think that the conditional is true at every nomologically possible world. It is certainly inapt to call such a position causally reductionist, but I'll leave the definition as it is. My goal is not to give a neat and intuitive taxonomy of theories of causation; my goal is just to subdivide logical space into three regions and argue that two of them are false. If you're not happy with the name I've given this area of logical space, feel free to substitute another.

[9] It seems as though most causal reductionists will be committed to something stronger, namely that, of nomic necessity, where C reduces to c and E reduces to e, c caused e iff C caused E. Even though, when a property Q supervenes upon a property P, we can only conclude that $\Box(\forall x(Px \Rightarrow Qx))$, and not $\Box(\forall x(Px \iff Qx))$, the view developed by Kim (1984b,a) is that, at any world at which C reduces to c, and E to e, what it is for C to cause E at that world is just for c to cause e at that world. Given this, it follows that, *at any world*, if C reduces to c at that world, and E reduces to e at that world, then c caused e iff C caused E. Anyone who accepts this stronger thesis will still count as a causal reductionist, in the sense that I am using the term here. However, accepting this stronger thesis is not necessary for being a causal reductionist, as I am using the term here.

My use of the terms *reduction* and *reductionism* here is slightly idiosyncratic. In the first place, many philosophers, taking cases like the reduction of water to H_2O as paradigmatic of reduction in general, believe that if A reduces to B, then A is identical to B. The causal reductionist is not a reductionist in this sense, since the high-level causal relation between C and E cannot be identical to the low-level causal relation between c and e. After all, C is distinct from c and E is distinct from e, and identical instances of the causal relation must relate identical events. While the causal reductionist does not take high-level causal relations to be *identical* to low-level causal relations, they generally think in some sense that the high-level causal relations *decompose into*,[10] or are *generated by*,[11] the low-level causal relations. They generally think that the facts in virtue of which C caused E just are the facts in virtue of which c caused e (and perhaps just this fact, that c caused e, itself).[12] Despite these terminological issues, I'm going to stick to *reduction*, since this is the word used by proponents of the view such as Kim and Price, as well as opponents of the view such as Menzies.

We might worry that not all high-level events have corresponding low-level events. For instance, we might worry that there is no low-level realizer of negative events such as Obama's failure to close Guantanamo.[13] Alternatively, if we think that, for instance, genetic origin is an essential property of persons, then we might worry that a single low-level event—the event of all the particles making up David's body having the fundamental properties that they do during his sneeze, *e.g.*—could correspond to multiple high-level events, since a particle-for-particle duplicate of a person could count as a different person in virtue of that particle-for-particle duplicate having a different genetic origin. In the actual world, this low-level event could realize *David's* sneeze. In some other world, it could realize somebody *else*'s sneeze, and not realize David's sneeze, since the low-level event in that world realizes the sneeze of somebody with a different genetic origin than David. If we think that reduction requires supervenience (E supervenes upon e iff it is impossible for E to fail to occur without e failing to occur), this would mean that the event of David's sneezing does not reduce to the event of the particles that compose David having the fundamental physical properties they

[10] See Price (1992).

[11] See Menzies (1988).

[12] Though, of course, they could accept the necessity of the conditional 'c caused $e \Rightarrow C$ caused E' without making any of these extra claims.

[13] Thanks to Brian Weatherson for this point.

do during his sneeze. So, we might worry, there is no event that realizes the high-level event of David's sneezing, even though David's sneezing caused me to say 'God bless you.'[14]

For better or worse, the causal reductionist is committed to the claim that every high-level event which enters into causal relations has some low-level realizer. Without this assumption, they cannot even formulate their view. The thesis I've labeled *Causal Reductionism* should, therefore, be taken to presuppose that there is some reduction relation between every causally-related high-level event and some low-level realizer. If this claim is false, then *Causal Reductionism* is false (or perhaps truth-valueless). In contrast, neither the causal eliminativist nor the causal emergentist is committed to the claim that every high-level event corresponds to some low-level event. It could be that negative high-level events occur in virtue of the fact that *no* low-level event of the appropriate kind occurs, for instance. Or it could be that which high-level event occurs at a time and place depends upon more than the low-level state of the world at that time and place.

Indeed, if one accepts that there are causally related coarsely-individuated events without any low-level realizers (I'll use the phrase 'low-level realizer' in such a way that, if e is low-level, then its low-level realizer is itself), then one has at hand a quick and easy argument for causal emergentism. For if there are coarsely-individuated events without any low-level realizers, then those coarsely-individuated events must be high-level. If those high-level events are causally related, then causal eliminativism is false. And, since those high-level events don't reduce to any low-level events, causal reductionism is false as well. So causal emergentism must be true.

This quick and easy argument is likely to be just as quickly and just as easily resisted. In the first place, the causal reductionist or the causal eliminativist may wish to deny the possibility of either absence causation or causal relations involving negative events, and thus deny the claim that Obama's failure to close Guantanamo caused the New York Times to write an editorial. They may similarly want to deny that the event of *David's* sneezing really caused me to say 'God bless you'. What *really* caused me to say 'God bless you', they may say, is that *somebody* sneezed. And this event does supervene upon the particles that compose David having the fundamental physical properties that they do. So I won't lean too heavily on this argument; however, for those who accept the claim that there are coarsely-individuated events without low-

[14] Thanks to David Kaplan and Calvin Normore for helpful conversation on this point—the example is due to Kaplan.

level realizers, they have at hand excellent reason to be causal emergentists even before getting to the arguments in §2.3.

Some causal reductionists might wish to formulate the thesis slightly differently than Kim. For instance, Huw Price writes that

> many of us think that there is something fundamental about microphysics. For one thing, we like to think that higher-level properties and relations obtain in virtue of fundamental physical properties and relations. For another, we like to think that big physical things and events are simply collections of little physical things and events. As a result, we are attracted to the idea that macroscopic causation is constituted by a lot of microscopic causation. One aspect of this intuition is that causal connections between temporally separated events decompose 'horizontally' into chains of more immediate causal connections; another is that these immediate connections, if not already primitive, decompose 'vertically' into a complex of microphysical causal relations. (Price, 1992, p. 511)

In Price's 'vertical' decomposition claim, he proposes that two causally-related high level events may each be constituted by, not a *single* low-level event, but rather *a complex* of low-level events, and what it is for the high level events to be causally related is just for a complex of causal relations to obtain between these low-level constituting events. These low-level constituting events differ from the high-level events along two dimensions. First, they are more finely-grained than the high-level events. Secondly, they are more *micro* than both the high-level events and the low-level realizers of those high-level events—they describe a smaller region of space-time. For instance, suppose that we have a gas with a certain temperature in a chamber with a certain volume. If we move one of the walls of the chamber so as to shrink the volume that the gas occupies, then the temperature of the gas will rise. The decrease in volume caused the increase in temperature. On Price's view, this causal relation decomposes vertically into a complex of causal relations between the particles making up the wall of the chamber and the particles composing the gas. As the wall is moved inward, the particles composing the wall collide with the particles composing the gas; since the particles in the wall are moving, they impart momentum to the particles composing the gas, thereby increasing those particles' velocities. And the gas's temperature is just a function of the velocities of the particles composing it. So as those velocities increase, the gas's temperature increases. Then, Price's thought is just this: what it is for the decrease in volume to cause the increase in temperature is just for a bunch of collisions to occur between the particles making up the wall of the chamber and the particles making up the gas. In each of these collisions, the movement of the particles in the wall cause the velocities of the particles they collide with to increase. So the high-level causal relation decomposes into a bunch of low-level causal relations.

How many of the low-level micro events which compose the low-level realizer of C must be causally related to the low-level micro events which compose the low-level realizers of E? And how many of the low-level micro events which compose the low-level realizers of E must they be related to? Is it enough that a single micro low-level event which composes the low-level realizer of C cause a single micro low-level event which composes the low-level realizer of E? Does it matter how large a part of the realizers of C and E these micro low-level events events make up? Difficult questions, all. It would be better if Price could avoid them entirely. Perhaps he can. A high-level event will be constituted by a complex of many low-level events. Take the mereological fusion of those events.[15] This mereological fusion will then just be the low-level realizer of the high-level event. Then, Price could simply say that two high-level events are causally related if the fusion of the complex of low-level events making them up are causally related. Such a view would be equivalent to the thesis I've called *Causal Reductionism* above. Some reductionists may want to resist this reformulation. On the one hand, they might think that some fusions are too spatiotemporally gerrymandered to enter into causal relations. On the other hand, they might worry about the fact that this reformulation appeals to the principle that a complex of events are causally related only if their mereological fusions are causally related. However, a reductionist might think that c_1 and c_2 can both cause e without their fusion causing e. I won't have much here to say about these worries. If the reader finds them persuasive, they are free to resist my reformulation. Nevertheless, unless Price says quite a bit more about how much or what kind of low-level micro causation is takes to have high-level macro causation, or what the relationship is between the low-level *micro* causal relations and the low-level *macro* causal relations, it will remain unclear whether his view entails *Causal Reductionism* or its negation, or whether it is consistent with both *Causal Reductionism* and its negation. In any case, whether we accept the reformulation or not, the argument I will provide in §2.3 will apply, *mutatis*

[15] I'm presupposing a theory of events according to which they are just regions of space-time, or properties of regions of space-time. If you think that events are facts, take the conjunction of these facts.

mutandis, to Price's version of causal reductionism as well. So it won't much matter, at the end of the day, whether I can squeeze Price into the Kimian framework. A Price-style reductionist will run into precisely the same troubles.

2.2.2 Causal Eliminativism

Causal reductionists accept that there *are* high-level causal relations, distinct from the low-level causal relations into which they decompose. They accept that Chris's smoking caused his cancer and that Zimbabwe's monetary policy caused its hyperinflation. They simply believe that these causal relations can be reduced to causal relations between low-level events. There is another position which denies that there are any high-level causal relations to be reduced. Call this position *causal eliminativism*. In general, causal eliminativists believe that the causal relations between low-level events are all the causal relations there are.[16]

There are two ways of being a causal eliminativist. Firstly, you could deny that there are any high-level events. That is, you could think that 'The Weimar Republic's monetary policy' and 'the Mark's hyperinflation' denote low-level events. Then, if you accept the truth of the causal claim 'The Weimar Republic's monetary policy caused the Mark's hyperinflation', you take the causal relation asserted by this claim to be a relation between two low-level events. Call this position *strong causal eliminativism*—strong because it does away with both high-level causal relations and high-level events.

Secondly, you could retain high-level events, and simply deny that these high-level events are causally related. Call this position *weak causal eliminativism*—weak because it only does away with high-level causal relations, retaining the high-level events. It may be thought that weak causal eliminativism provides its own reductio. To accept that there are high-level events yet deny that these events enter into causal relations is to deny that carbon emissions caused global temperatures to rise. It is to deny that asbestos ever caused mesothelioma and that oxidation ever caused rust. Intelligent people ought not deny such things, and certainly not on account of philosophical considerations alone. Unfortunately, weak causal eliminativism cannot be dismissed so quickly. Sophisticated weak eliminativists will accept the assertibility of the English sentence 'Chris's smoking caused his cancer', while denying that this sentence is used to assert the existence of a causal relation between the smoking and the cancer. The

[16] Assuming that they think that there are low-level causal relations. Denying the existence of both high- and low-level causation is another way of being a causal eliminativist. Russell (1912) provides an example of this brand of causal eliminativism.

most developed version of this account that I am aware of comes from Strevens (2008). According to Strevens,

> claims of the form *c was a cause of e*...do not assert the existence of a raw metaphysical causal relation between two events *c* and *e*; rather, they are causal-explanatory claims that assert that *c* is a part of the causal explanation for *e*.[17]

Strevens is here picking up on a suggestion of Davidson (1967)'s: namely that, in sentences like 'The collapse was caused, not by the fact that the bolt gave way, but rather by the fact that it gave way so suddenly and unexpectedly', the verb 'caused'

> is not the 'caused' of straightforward singular causal statements, but is best expressed by the words 'causally explains.'[18]

What Davidson suggests as a fix for a few recalcitrant sentences, Strevens adopts for all, or at least most, of our everyday causal claims. I'll call this thesis, that causal claims are causal explanatory claims, the 'Davidson-Strevens thesis'.

There are two ways of understanding the Davidson-Strevens thesis. We could understand it as a semantic claim: the literal content of 'C caused E' is that C is a part of a causal explanation of E—*i.e.*, the semantic value of 'caused' in a sentence 'C caused E' is a function that takes the value *true* iff the low-level realizer of C is causally related to the low-level realizer of E and the fact that C occurred explains the fact that E occurred. Alternatively, we could understand it as a claim about the pragmatics of causal claims: in ordinary contexts at least, to say 'C caused E' is to pragmatically implicate that the low-level realizer of C is causally related to the low-level realizer of E and that the occurrence of C explains the occurrence of E, even though the literal content of 'C caused E' is just that C and E are causally related, and not that the occurrence of C is explanatorily relevant to the occurrence of E. This pragmatic story, wedded with weak causal eliminativism, entails that most causal claims are used to implicate true things despite being strictly speaking false. (This is not as odd as it may seem. On many semantic theories, we often use false sentences to pragmatically implicate something true—*e.g.*, 'You can't get there from here' and 'I have nothing to wear.')

[17] Strevens 2008, p. 4. See also Strevens 2008, ch. 6.
[18] Davidson 1967, pp. 161–2.

This sophisticated form of weak causal eliminativism is, therefore, not merely a metaphysical claim; it is also a claim about the semantics (or the pragmatics) of causal talk. Of course, the view that causal claims are (or pragmatically implicate) causal explanatory claims could be combined with strong causal eliminativism as well. Indeed, as a purely semantic (or pragmatic) thesis, it is compatible with any of the positions I am considering here. However, for the weak causal eliminativist, this thesis, or something like it, is necessary to avoid the absurd conclusion that the sentence 'carbon emissions caused climate change' is unassertable.

2.2.3 Causal Emergentism

The remaining logical space is occupied by a character I will dub the *causal emergentist*. Many of the so-called British emergentists embraced emergentism about laws of nature. They held that the higher-level laws of nature governing complex phenomena such as chemical reactions or the biological activity of living organisms did not reduce to the fundamental laws of physics.[19] To have a name, call this kind of emergentism *nomic emergentism*. Just as the nomic emergentist thinks that there are higher-level laws of nature which cannot be reduced to lower-level laws of nature, the *causal emergenist* believes that there are causal relations between high-level events which cannot be reduced to causal relations between their low-level realizers. They agree with the causal reductionist, against the causal eliminativist, that there is high-level causation. However, they reject the material conditional

$$c \text{ caused } e \qquad \Rightarrow \qquad C \text{ caused } E$$

(where c and e are the low-level realizers of C and E, respectively) for some C and E.

Note that, in rejecting this conditional, the causal emergentist needn't commit themselves to the claim that there *are* low-level realizers of high-level events. Rejecting the idea that some high-level events reduce to low-level events while maintaining that those high-level events are causally related is just one way of being a causal emergentist.

Causal emergentism does not entail nomic emergentism. Moreover, you could be a causal emergentist without denying that causal relations between high-level events can be reduced to *some* low-level facts. In order to count as a causal emergentist, it is enough to claim that the low-level facts in virtue of which high-level events are causally related are not exclusively low-level *causal* facts about the realizers of those

[19] See McLaughlin (1992).

high-level events. You *may*, of course, accept a higher-octane version of emergentism, according to which causal relations between high-level events are not determined by or in principle predictable on the basis of the low-level state of the world and the low-level laws of nature; but there's nothing in the thesis of causal emergentism demanding such an extravagant claim. Causal emergentists can—and, by the way, should—accept that the behavior of every physical object is entirely determined by the fundamental physical state of the world and the fundamental physical laws.[20]

To see this, return to Lewis's 1973 account of causation. Lewis accepted that the high-level causal facts were determined by the low-level facts.[21] Nevertheless, his original counterfactual account of causation entails causal emergentism. On that account, recall, an event e causally depends upon a distinct event c iff were c to fail to occur, e would fail to occur as well. Causation is the ancestral of causal dependence. This account is in need of a theory of events, and the conditions under which they do and do not occur. This was provided by Lewis's 1986 paper *Events*. There, Lewis claims that an event is a property of a space-time region. Since, for Lewis, properties are just classes of individuals at worlds, a property of a space-time region is just a class of space-time regions at worlds. A space-time region at a world has the property just in case it belongs to the class. For an event e to occur at a world is for one of its members to exist at that world. Events can be classified by their essences, the conditions necessary for them to occur—*i.e.*, the conditions met by all of their members.

Lewis tells us that some events *imply* other events, where an event e implies an event f iff, necessarily, if e occurs in region R, then f occurs in region R. He illustrates this kind of implication with the following example: John's saying 'hello' loudly implies John's saying 'hello'. Though he expresses some hesitation at taking these two events to be distinct, Lewis decides that they must differ, since they differ causally. John's saying 'hello' caused Fred to greet him; whereas John's saying 'hello' loudly did not. (Fred would still have greeted John even if he had said 'hello' at a normal volume.) Similarly, John's saying 'hello' loudly was caused by John's being a bit drunk; whereas John's saying 'hello' was not caused by John's being a bit drunk. (Had John

[20] Along, perhaps, with some brute facts about the outcome of tychistic chancy processes, like the collapse of the wave function on some interpretations of Quantum Mechanics.

[21] Given Lewis's physicalism, the low-level facts will specify a unique possible world. The laws at this world come along for free—they are the generalizations of the axiomatic system of truths which strikes the best balance of simplicity, strength, and fit (see Lewis 1983, 1994a). The semantics for counterfactuals given in Lewis (1973b, 1979b) then provides the truth conditions for causal relations both high and low.

not been drunk, he would not have said 'hello' loudly; however, he would still have said 'hello'.) LEWIS describes the relation between these two events thusly:

> We have, so to speak, a more and a less detailed version of what happens in a region. Both are occurrent events. The more detailed version has a richer essence; the otherworldly regions included in it are fewer and less varied...The more detailed version is one, but only one, of the ways in which the less detailed version could have occurred.[22]

The relationship between high-level events and their low-level realizers is analogous to the relationship between John's saying 'hello' and John's saying 'hello' loudly. Low-level events imply the high-level events they realize. The low-level event is a more detailed version of what happens in a region; it has a richer essense than the high-level event it realizes. The low-level realizer is one, but only one, of the ways in which the high-level event could have occurred. For this reason, just as John's saying 'hello' loudly can be caused by things that John's saying 'hello' is not, a low-level event can be caused by things which the high-level event it realizes is not. That's because the closest possible world at which the cause fails to occur could be a world at which the low-level realizer fails to occur, yet the high-level event it realizes doesn't fail to occur.

In addition, two low-level events c and e can be causally related without the high-level events they realize, C and E, being causally related. For the closest world at which c fails to occur could be a world at which e fails to occur without the closest world at which C fails to occur being a world at which E fails to occur. For instance, suppose that the modal profiles of c, e, C, and E are as shown in figure 1. There, the counterfactual $\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$ is true, while the counterfactual $\neg O(C) \mathrel{\Box\!\!\rightarrow} \neg O(E)$ is false. Suppose further that there is no sequence of events $D_1 ... D_N$ such that $\neg O(C) \mathrel{\Box\!\!\rightarrow} \neg O(D_1) \wedge ... \wedge \neg O(D_N) \mathrel{\Box\!\!\rightarrow} \neg O(E)$. Then, LEWIS's 1973 account will tell us that c caused e even though C didn't cause E. So it is not nomically necessary that C cause E whenever c causes e.

Due to the universality of fundamental physical forces like gravitation and electromagnetism, examples with this structure are not difficult to come by.[23] The moon

[22] LEWIS 1986b, p. 257.

[23] In the body, I'm going to pretend that the fundamental physical state of the world and the fundamental laws of nature are roughly what we thought that they were about a century and a half ago. I don't believe, however, that any of the arguments depend upon this assumption. Under more realistic assumptions, extra provisos about past light cones and the like will be needed. When appropriate, I'll make the necessary revisions in the footnotes.

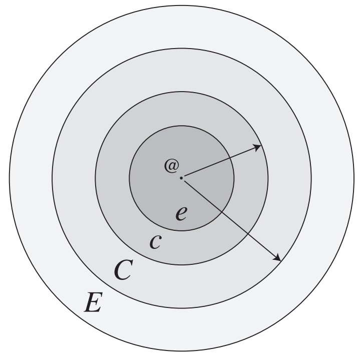


Figure 1: In the diagram, simiarity is represented with distance. All and only the worlds inside the innermost circle are worlds at which e occurs; all and only those inside the next largest circle are those at which c occurs; and so on. The closest world to the actual world, @, at which c does not occur is a world at which e does not occur. Yet the closest world at which C does not occur is not a world at which E does not occur.

wanes, and the window shatters. The waning of the moon is microphysically realized by the fundamental particles which constitute the moon having certain masses and charges, and taking certain precise trajectories over a certain period of time. Similarly, the shattering of the window is microphysically realized by the individual particles which constitute the window having certain masses and charges and taking certain precise trajectories over a certain period of time. Both of these are incredibly fragile events. Because their essences are so rich, it is very easy for them to fail to occur—if any of the particles in the window were to have an ever-so-slightly different mass or charge or take an ever-so-slightly different trajectory during the time period, then the low-level realizer of the shattering would fail to occur; and likewise for the masses, charges, and trajectories of the particles in the moon and the low-level realizer of the moon's waning. Thus, had the low-level realizer of the moon's waning failed to occur, at least one of the particles which constitute the moon would have had a slightly different mass or charge or taken a slightly different trajectory. However, if any of those particles had had a slightly different mass or charge or taken slightly different trajectory, then it would have exerted a slightly different gravitational or electromagnetic force on all of the particles which realize the window's shattering. Since the future trajectories of the particles in the window are entirely determined by the resultant of the forces acting upon them, they would have taken slightly different trajectories, had any of the gravitational or electromagnetic forces acting upon them been slightly different. So the low-level realizer of the window's shattering would have failed to occur. So the low-level realizer of the moon's waning caused the low-level realizer of the window's shattering, on Lewis's 1973 account. But the moon's waning did not cause the window's shattering, since, had the moon not waned, the window would still have shattered. (It would have shattered in an ever-so-slightly different *way*, but it would have shattered all the same.) Assuming that there is no intermediate event which depends upon the moon's waning and upon which the shattering depends, it follows on the 1973 account that the moon's waning did not cause the window's shattering.

So, on Lewis's 1973 counterfactual account of causation, if you look at low-level events, then you see one causal structure. If you look at high-level events, you see a different causal structure. And which causal structure you see depends upon which high-level events you're looking at. There is one network of causes and effects leading into and out of John's saying 'hello' and another network of causes and effects leading into and out of John's saying 'hello' loudly.

Interestingly, this aspect of the counterfactual account goes away in Lewis's 2000 revision. On the revised account, recall, an *alteration* of an event is just a variation of the time or manner of the event's occurrence, whether that variation leads to the event failing to occur, or merely occurring in a slightly different manner, or at a slightly different time. An event c causally influences a distinct event e iff there is a substantial range of not-too-distant alterations of c, $c_1 ... c_N$, and a range of alterations of e, $e_1 ... e_N$ (at least some of which differ), such that the c_i counterfactually pattern with the e_i—that is to say, had c_i occurred, e_i would have occurred, for all i. Causation is the ancestral, or the transitive closure, of causal influence.

On this account, no longer can John's saying 'hello' loudly be caused by something which does not also cause John's saying 'hello'. If not-too-distant alterations of John's being drunk counterfactually pattern with alterations of John's saying 'hello' loudly, then they will also counterfactually pattern with alterations of John's saying 'hello'. That's because an alteration of John's saying 'hello' loudly *just is* an alteration of John's saying 'hello'. A variation in the time or manner of the event of John's saying 'hello' loudly just is a variation in the time or manner of the event of John's saying 'hello.' This variation might be a variation which makes it the case that John's saying 'hello' loudly doesn't occur, while his saying 'hello' still does occur, but that doesn't matter, on the 2000 account. So long as alterations in John's being drunk counterfactually pattern with alterations of John's saying 'hello', it doesn't matter whether those alterations are ones in which the saying 'hello' occurs in a different manner, or at a different time, or whether they are ones in which it doesn't occur at all.

For similar reasons, if not-too-distant alterations of the low-level realizer of the moon's waning counterfactually pattern with alterations of the low-level realizer of the window's shattering, then not-too-distant alterations of the moon's waning will counterfactually pattern with alterations of the window's shattering. That's because not-too-distant alterations of the low-level realizer of the moon's waning *just are* not-too-distant alterations of the moon's waning; and alterations of the low-level realizer of the window's shattering *just are* alterations of the window's shattering. So, if a low-level event c influences another low-level event e, then the high-level event C which c realizes must influence the high-level event E which e realizes—for *any* high-level events C and E which c and e realize. So, if c influences e, then C influences E.[24]

This is just a claim about influence. Causation, however, is the ancestral of influence. If c causes e by influencing it directly, then, since—as we just saw—if c influences e, then C influences E, we can conclude that C influences E, and therefore, that C causes E. If, on the other hand, c causes e by being connected to it by a chain of influence $c \to d_1 \to d_2 \to ... \to d_N \to e$, then we know that not-too-distant alterations of c counterfactually pattern with alterations of d_1. But not-too-distant alterations of c just are not-too-distant alterations of C, so not-too-distant alterations of C must also counterfactually pattern with alterations of d_1. And, since d_N influences e, not-too-distant alterations of d_N counterfactually pattern with alterations of e. But alterations of e just are alterations of E, so not-too-distant alterations of d_N must also counterfactually pattern with alterations of E. So there is a chain of causal influence running from C to E *via* $d_1, d_2, ..., d_N$. Since causation is the ancestral of causal influence, C caused E. So, whether c caused e by influencing it directly or by being linked to e by a chain of influence, if c caused e, then C caused E. So the 2000

[24] We might worry about the requirement that the range of alterations of C be *substantial*. It could be that what counts as a substantial range of alterations for c does not count as a substantial range of alterations for C. Perhaps. But even so, the range of alterations of c which counterfactually pattern with alterations of e will be at least a subset of a substantial range of alterations of C; and if a set of alterations counterfactually pattern with a range of alterations of e, then any superset will also counterfactually pattern with a range of alterations of e, since Lewis allows arbitrarily many of the alterations of e to be identical, so long as some of them differ. Gordon Belot (p.c.) has pointed out that this renders the requirement that the range of alterations of c be *substantial* an empty one; whenever two alterations of c counterfactually pattern with e, a proper superset of these alterations will as well. I'm more inclined to see this as an objection to Lewis's formulation of his definition of influence than as an objection to my interpretation of Lewis, especially given that Lewis in several places (*e.g.*, p. 188) cites two alterations of an event which counterfactually pattern with two distinct alterations of another event, and straightaway concludes that the first event influences the second.

account entails causal reductionism.[25]

Before moving on, let me stress again that *causal* emergentism, unlike *nomic* emergentism, does not entail that there are irreducible higher-level laws or even that high-level causal relations are irreducible. Causal emergentists can, like Lewis (1973), take high-level causal relations to be entirely reducible to fundamental physical facts without taking them to be entirely reducible to low-level *causal* facts. So when Kim writes that

> macro-causality...must be viewed not as something basic and fundamental but as something that is reducible to, and explainable in terms of, more fundamental causal processes[26]

he poses a false dichotomy. We need not choose between claiming that high-level causal relations are basic and fundamental and claiming that they are reducible to low-level causal relations. We can insist that the high-level causal relations are entirely reducible to low-level facts without claiming that they are entirely reducible to low-level *causal* facts.

2.2.4 A Taxonomy

In summary, we can categorize these three positions according to whether they agree with the following claims.

High-Level Causation: There are high-level causal relations.

High-Level Events: There are high-level events.

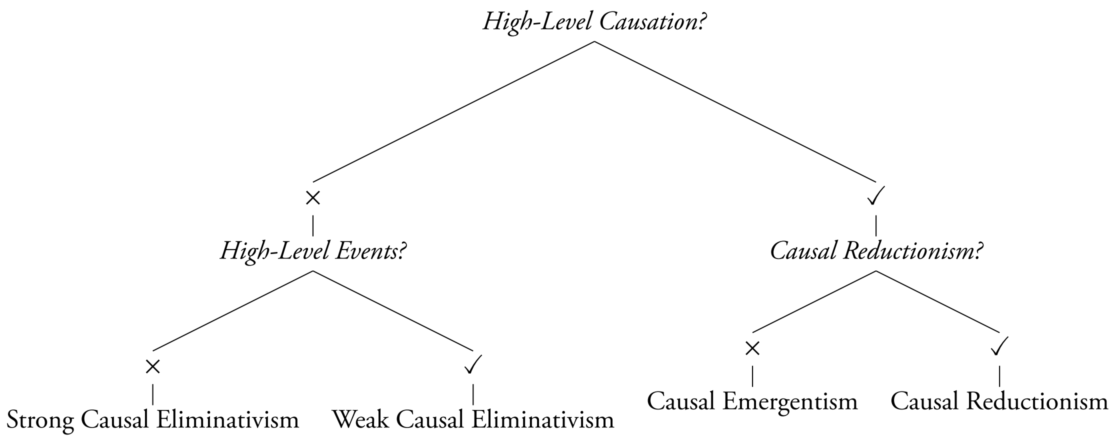
Causal Reductionism: Of nomic necessity, where C reduces to c and E reduces to e, if c caused e, then C caused E.[27]

[25] Brian Weatherson (p.c.) worries that Lewis (2000) might not count as a causal reductionist, since he might not accept that all high-level events that are causally related reduce to low-level events. The *implication* relation gives us some relation between the high-level and the low-level events; however, it may be that, on Lewis's view, some high-level events that enter into causal relations—like, *e.g.*, absences (though these are not events in Lewis's terminology, they are events in mine)—are not implied by, or otherwise appropriately related to, any low-level event. This is an interesting exegetical question, but one that I'll set aside for now. Even if Lewis (2000) is not technically a causal reductionist, he is still much *more* reductionist than he was in 1973; and more reductionist than the arguments in §2.3 will allow.

[26] Kim (1984b, p. 51).

[27] We should understand this thesis in such a way that it is false if there is no appropriate reduction relation between high- and low-level events.

If you reject *High-Level Causation*, then you are a causal eliminativist. If you additionally reject *High-Level Events*, then you are a strong causal eliminativist. If you additionally accept *High-Level Events*, then you are a weak causal eliminativist. If you accept *High-Level Causation* and accept *Causal Reductionism*, then you are a causal reductionist. If you accept *High-Level Causation* but deny *Causal Reductionism*, then you are a causal emergentist.



2.3 A Defense of Causal Emergentism

In this section, I'm going to argue that both causal reductionism and causal eliminativism are untenable. In brief, the problem I will pose for the causal reductionist is that high-level causal relations appear to be relatively sparse, whereas low-level causal relations appear to be relatively abundant. These two claims are in tension with the thesis that high-level events are causally related if the low-level events which realize them are causally related. The same objection applies, *mutatis mutandis*, to the strong causal eliminativist. The tension between the apparent abundance of low-level causation and the apparent sparsity of high-level causation has been noted before,[28] but it has been thought that the tension can be massaged and mitigated in various ways. In this section, I'm going to try to make trouble for those mitigating strategies. In the course of stirring up that trouble, I'll provide an independent argument against what I earlier called the Davidson-Strevens thesis—that causal claims are (or implicate) causal explanatory claims. Since weak causal eliminativism is only as plausible as the

[28] See Russell (1912), Field (2003), and Price & Corry (2007).

Davidson-Strevens thesis—without this thesis, the weak eliminativist is forced to say that sentences like 'carbon emissions caused climate change' are unassertable—an argument against the Davidson-Strevens thesis is an argument against weak causal eliminativism as well. I will therefore counsel a rejection of causal reductionism and causal eliminativism, and an acceptance of causal emergentism.

More carefully, the problem for causal reductionism and strong causal eliminativism is this: the following three claims form an inconsistent set.

10. A coarsely-individuated event C caused another coarsely-individuated event E if C's low-level realizer caused E's low-level realizer.

11. The orbit of Gliese 163c did not cause stagflation.

12. The low-level realizer of the orbit of Gliese 163c did cause the low-level realizer of stagflation.

(Gliese 163c is a potentially habitable planet, approximately 49 light years from Earth, first discovered in 2012. Throughout, let's take the phrase 'the orbit of Gliese 163c' to refer to Gliese 163c's taking a single trip around its star, beginning at the start of Earth year 1900.) In 10, I'm going to understand the phrase 'low-level realizer' in such a way that, if C is low-level, then C's low-level realizer is itself. Then, the strong causal eliminativists thinks that 10 says that if C caused E, then C caused E. Since this is a truth of logic, the strong causal eliminativist is committed to 10. The causal reductionist takes 10 to just be a statement of their view, so they are committed to it as well. I will contend that both 11 and 12 are true. Since 11 and 12 are true, 10 is false. From the falsehood of 10, it follows that both causal reductionism and strong causal eliminativism are false.

11 really should not need defense. It is a Moorean fact if any is. Perhaps there are possible worlds in which the orbit of Gliese 163c led to a decrease in the growth rate of production, an increase in the growth rate of prices, and above average unemployment in the United States during the 1970's. But ours is not a world like that. Imagine seriously suggesting the negation of 11 to an economist. We can put in all the qualifications we like: 'I don't mean to suggest that Gliese 163c's orbit was an *interesting* or *salient* cause of stagflation. But it was, nevertheless, one of the causal factors leading to stagflation.' Even with these qualifications, we should be embarrassed to say such a thing. The precise duration and severity of stagflation in the U.S. is entirely determined by three factors: the number of goods produced and consumed in the U.S.,

the price of each of those goods, and the percentage of the workforce employed. And not a single job in the U.S. economy would have been saved, not a single firm's production, nor any consumer's consumption, would have been any higher, nor a single price any lower, had the orbit of Gliese 163c occurred differently or not at all. And that's not because this is a recherché case of preemption, where something else would have done the work of Gliese 163c's orbit had Gliese 163c not orbited. The right thing to say about Gliese 163c's orbit is that it simply *wasn't* a causal factor for stagflation, however slight, however uninteresting, however negligible. (But I will have more to say about this below.)

In support of 12: every major theory of causation in good standing gets the result that the low-level realizer of Gliese 163c's orbit caused the low-level realizer of stagflation. Start with the counterfactual account. As we saw above with respect to the moon's waning and the window's shattering, the low-level realizer of an event like stagflation or the orbit of Gliese 163c consists in certain fundamental physical particles having certain masses and charges and taking certain precise trajectories over a certain period of time. These events are incredibly fragile—if the masses, charges, or trajectories of those fundamental particles were to differ in the slightest, those low-level events would fail to occur. Apply the counterfactual test for causation: were the microphysical realizer of Gliese 163c's orbit to fail to occur, at least one of the particles which realize Gliese 163c's orbit would have differed, however slightly, in its mass, charge, or trajectory; it would therefore have exerted a slightly different gravitational or electromagnetic force upon all of the particles which realize stagflation. Since the trajectories of those particles are completely determined by the resultant of the forces acting upon them, the trajectories of those particles would have differed, however slightly. So the low-level realizer of stagflation would have failed to occur. So the low-level realizer of stagflation counterfactually depends upon the low-level realizer of Gliese 163c's orbit. No philosopher in the counterfactual tradition claims that non-backtracking counterfactual dependence between distinct events like this is necessary for causation. But almost all accept that non-backtracking counterfactual dependence between distinct events is a sufficient condition for causation.[29] And that

[29] Besides the non-backtracking and distinct event qualifications, we should also add that the counterfactuals relate intrinsic properties of the events in question. If we're allowed to appeal to mere Cambridge properties, then the counterfactual test would imply that a distant supernova could cause me to become such that a supernova has recently occurred. However, many find this result unpalatable. (Not the least because it seems to involve action-at-a-distance.) See Maudlin (2011).

is all we need here in order to conclude that the low-level realizer of Gliese 163c's orbit caused the low-level realizer of stagflation.

Consider instead a regularity account of causation. On Mackie (1965)'s account, for instance, *c* is a cause of *e* iff *c* is a part of a minimally sufficient condition for *e* which actually obtains. And the low-level realizer of Gliese 163c's orbit, together with the simultaneous state of the rest of the universe, is a minimally sufficient condition for the low-level realizer of stagflation. Given determinism, the entire state of the universe at a time, together with the laws of nature, is both necessary and sufficient for the state of universe at every other moment—and, in particular, for the part of the universe which realizes stagflation.[30]

Or consider a probabilistic account of causation. According to the most plausible versions of those accounts, *c* caused *e* just in case *c* changes the probability of *e* in a causally homogenous background context. A causally homogenous background context is given by all of the causes of *e*, except *c* and events caused by *c* (if *c* is indeed a cause of *e*). So, given the universality of the fundamental forces, the entire state of the rest of the universe at a time, other than the realizer of the orbit of Gliese 163c, constitutes a causally homogenous background context for the microphysical realizer of stagflation—call that background context 'K'. K does not entail that the microphysical realizer of stagflation occurs, so, given some plausible assumptions, the probability of the microphysical realizer of stagflation, given K, should be less than 1. However, assuming determinism, the probability of the microphysical realizer of stagflation, given K and the realizer of the orbit of Gliese 163c, will be 1. So, on the probabilistic account, the realizer of the orbit of Gliese 163c caused the realizer of stagflation.

Or consider a process theory of causality, like that of Dowe (2000) or Salmon (1994). On Dowe's account, a low-level event *c* caused another low-level event *e* if, roughly, *c* and *e* are connected by a series of causal processes whose intersections constitute causal interactions. A causal process is just the world line of an object which possesses a conserved quantity, and a causal interaction is an intersection of world lines that involves the exchange of a conserved quantity. Photons collide with the particles

[30] On certain interpretations of Quantum Mechanics, the entire state of the universe at one time only determines a probability distribution over the state of the universe at future times. In that case, the entire state of the universe at one time, together with facts about the outcome of tychistic chancy processes—the collapse of the wavefunction—will constitute a minimally sufficient condition for the state of the universe at future times.

which make up the realizer of Gliese 163c's orbit. These collisions constitute causal interactions; in the collisions, there is an exchange of momentum between the particles and the photons. Some of these photons make their way to Earth; some of those collide with some of the particles which make up the low-level realizer of stagflation. These collisions also count as causal interactions; momentum is exchanged between the photons and the particles. So there is a causal process leading from the low-level realizer of Gliese 163c's orbit to the low-level realizer of stagflation. So, on Dowe's process theory, the low-level realizer of Gliese 163c's orbit caused the low-level realizer of stagflation. (Similar remarks apply to Salmon's process theory.)

And since an in-principle intervention upon the state of the low-level realizer of Gliese 163c's orbit would bring about a change in the state of the low-level realizer of stagflation (for the very reason that the counterfactual 'had the low-level realizer of Gliese 163c's orbit not occurred, the low-level realizer of stagflation would not have occurred' comes out true) the manipulationist account of Woodward (2003) will rule the realizer of Gliese 163c's orbit a cause of the realizer of stagflation.

Every major philosophical theory of causality in good standing implies that the low-level realizer of Gliese 163c's orbit and the low-level realizer of stagflation are causally related. I conclude that they are causally related. So I2 is true. Since I1 is also true, causal reductionism and strong causal eliminativism are both false.

Some respond to these kinds of considerations by rejecting I1. They accept I2, and this, together with their reductionism or strong eliminativism, entails that the orbit of Gliese 163c caused stagflation. So they countenance the counterintuitive causal relationship, but insist that they can explain away the intuitive appeal of I1 with various semantic or pragmatic theses about our causal talk. Here is Alyssa Ney (2009) endorsing this general strategy:

> Although it is true that the foundationalist picture of causation under consideration[31] has the consequence that due to the multitude of fundamental causal interactions, there are many, many factors of causal influence for every event we might consider, we ordinarily want to single out at most a few as 'the causes' of an event, in our ordinary causal assertions. (Ney, 2009, p. 741)

[31] While she doesn't say enough for me to precisely locate her in the taxonomy of 2.2.4, the view of causation Ney (2009) calls 'foundationalist' is at least committed to the disjunction of causal reductionism and causal eliminativism.

For the remainder of this chapter, therefore, I will be considering the plausible semantic or pragmatic theses which it has been supposed are capable of successfully explaining away the intuitive appeal of 11, and arguing that none of them meet with success. Of course, it is worth emphasizing that there will *always* be some pragmatic story the reductionist or eliminativist can tell which will be capable of getting all the data right. Suppose, for instance, that you have produced a emergentist theory of causation, call it 'T_e', which aligns perfectly with our characteristic causal judgments in every case; it says that two events are causally related when and only when we judge the corresponding causal claim to be true. Suppose then that a reductionist puts forward a theory of causation, call it 'T_r', which does a much worse job aligning with our characteristic causal judgments; it frequently asserts the existence of a causal relationship when we judge there to be none. Such a reductionist could always put forward the following semantic/pragmatic thesis: a sentence '*c* caused *e*' is true/assertible iff T_e claims that there is a causal relationship between c and e. This will be capable of capturing all of the data iff T_e was capable of capturing all of the data. But, I take it, such an account wears its implausibility on its sleeve. It's not enough to provide some *ad hoc* pragmatic theory of which causal claims are true/assertible. It must additionally *make sense*, given the theory of what the causal relation is and given general features of language, that we would talk in accord with the semantic/pragmatic theory. The two theses I will consider below—*viz.*, that we choose to ignore the causes which have negligible influence, and that we choose to ignore the causes which don't provide good explanations—both meet this minimal standard of adequacy. It is a general feature of language that we neglect the negligible. And, given that we often look to causal relations in order to provide explanations, it makes good sense that we would only focus on those that succeed in providing such explanations.

Turning now to those strategies: in the first place, a causal reductionist or a strong causal eliminativist may wish to suggest that, while Gliese 163c's orbit *did* cause stagflation, the causal influence that Gliese 163c's orbit had on stagflation was so minute that, for all practical purposes, it can be ignored. For this reason, we hesitate to accept 11. For instance, Lewis recognizes that his 2000 account commits him to saying that

> ...almost everything that precedes an event will be counted among its causes. By the law of universal gravitation, a distant planet makes some minute difference to the trajectory of Suzy's rock, thereby making a tiny

> difference to the shattering of the bottle...we open the gate to a flood of spurious causes.[32]

His response to this worry is that we are justified in ignoring these so-called spurious causes[33] on the grounds that their influence will be negligible:

> Well—these differences made by spurious causes are negligible, so surely we are entitled to neglect them.[34]

Here, Lewis diagnoses the oddity of 11 by appeal to a pragmatic thesis about which causes are most felicitously cited in a causal claim. The thesis, which I'll call 'the Lewis thesis', is that the causes with the most influence are most felicitously cited in a causal claim. Importantly, the Lewis thesis is a thesis about *relative*, not *absolute*, influence. It doesn't say that events with tiny influence will not be felicitously cited as causes; if *all* of an event's causes have tiny influence, then it could be that a cause with a tiny amount of influence is most felicitously cited in a causal claim, so long as its influence is greater than that of all of the other causes. Similarly, the Lewis thesis doesn't tell us that causes with large influence will always be appropriately cited in a causal claim; if there is another cause with still larger influence, it could be infelicitous to cite even a very influential cause.

The Lewis thesis tells us that 11 appears true because, while the orbit of Gliese 163c *did* cause stagflation, in ordinary linguistic contexts, we ignore influences as negligible as the orbit of Gliese 163c—just as, for instance, in ordinary linguistic contexts, the

[32] Lewis 2000, p. 188, Lewis 2004, p. 89.

[33] Lewis calls these causes "spurious", but he doesn't really mean it. On his 2000 account, they are genuine causes; they are merely appropriately treated as spurious in most contexts—but this is a matter of the pragmatics of causal talk, and not the metaphysics of causation. There seems to be some confusion about this in the literature. For instance, Paul & Hall (2013) claim that Lewis's 2000 view is that "*C* influences *E* just in case *E* counterfactually covaries with *C to a sufficient extent*" (p. 17, emphasis added). *Pace* Paul & Hall, the requirement that the counterfactual covariation be sufficiently large is not a part of the metaphysical account of influence; it is rather a pragmatic thesis about our causal talk. Lewis obscures this point considerably by talking about "spurious" causes and saying things like "Billy's throw does not influence the shattering" (Lewis 2004, p. 92), when what he really means is 'Billy's throw (near enough) does not influence the shattering.' (Of course, given Lewis's pragmatic thesis, he is within his linguistic rights to say these things.)

[34] Lewis 2000, p. 189. Interestingly, the quoted sentence changes between the 2000 *Journal of Philosophy* version and the 2004 version in *Causation and Counterfactuals*. While nothing else in the surrounding text changes, this sentence, which ends with a period in the 2000 version, ends with a question mark in the 2004 version.

sentence 'there's nothing in the fridge' is assertible, even though the sentence is false so long as there is air and dust in the fridge.[35]

As I said before, I don't think that the right thing to say about Gliese 163c's orbit is that it makes a *negligible* difference to stagflation. Given that not a single job would be saved, not a single firm's production would be higher, and not a single price would be lower without Gliese 163c's orbit—and given that these factors completely determine the precise duration and severity of stagflation—the right thing to say is that Gliese 163c's orbit didn't make *any* difference to stagflation, however negligible. But put that point aside. There's a bigger worry. The worry is that, often enough, seemingly spurious causes have quite a large influence on their putative effects—even larger influence than the apparently genuine causes.

Suppose that Sabeen tells you truthfully that she plans to slip a fatal poison into Stephanie's drink. You are unable to warn Stephanie, and you don't know how to neutralize the poison, but you do have on you a powerful anesthetic which will numb and immobilize Stephanie, making her death far less painful. You pour the anesthetic into Stephanie's drink. She drinks, and dies quickly and painlessly. According to Lewis's 2000 account, your pouring the anesthetic into Stephanie's drink caused Stephanie's death—since not-too-distant alterations in your pouring of the anesthetic counterfactually pattern with alterations in Stephanie's death; had you not poured, the death would have occurred in a different manner. However, the influence that it had on the death is by no means negligible. Had you not given Stephanie the anesthetic, she would have died a much more painful death. There would have been writhing and cursing and gnashing of teeth. Moreover, the influence of the fatal poison is comparatively small. Given the presence of the immobilizing anesthetic, alterations in the pouring of the poison counterfactually pattern with comparatively minor alterations of Stephanie's breathing, heartbeat, and other metabolic functions. (In fact, we can stipulate that, if the lethal poison had not killed Stephanie, then the anesthetic *would* have eventually prevented her from breathing, causing her to die shortly thereafter. Then, not pouring the poison would only slightly delay the death.) Nevertheless, we are loath to countenance your pouring of the anesthetic into Stephanie's drink as a cause of her death. We judge the sentence

13. # Your pouring the anesthetic into Stephanie's drink caused her to die.

[35] This is Lewis's diagnosis of sentences like 'there's nothing in the fridge.' One could, of course, think that the quantifier in this sentence is restricted to foodstuffs, rendering the sentence just plain old true.

to be in some deep sense inappropriate; and we judge the sentence

14. Sabeen's pouring the lethal poison caused Stephanie to die.

to be in some deep sense appropriate. It is my view that 13 is not only inappropriate, but false, but everybody should be able to agree that it is infelicitous to utter 13. The problem is that the Lewis thesis predicts that 13 should be felicitous, or at least *as* felicitous as 14 (if not more so). This prediction is not borne out; which gives us strong reason to reject the Lewis thesis.[36] Of course, no pragmatic thesis is ever decisively refuted; we can always wheel in other pragmatic considerations to explain away any recalcitrant data. However, explaining the infelicity of sentences like 13 was precisely the kind of work the Lewis thesis was meant to accomplish. If the thesis cannot explain the infelicity of 13, it is not doing its work well.

In the foregoing argument, I supposed that, if minor alterations of an event c_1 counterfactually pattern with large macroscopic alterations in an event e, and minor alterations of an event c_2 counterfactually patterns with macroscopically undetectable alterations in e, then c_1 influences e more than c_2 does. In conversation, several people have suggested that a causal reductionist or a strong causal eliminativist should deny this principle, and instead maintain that which influences are greater than others should be a context-sensitive matter. In the case presented above, for instance, they might suggest that, while the poison didn't make any macroscopically detectable difference to the event of Stephanie's death; it still made a difference to *whether it was a death*. Such a strategy would help in the present case, but it would fail in general. So understood, the Lewis thesis falls immediately to cases of preemption; alterations of neither Billy's nor Suzy's throw counterfactually pattern with alterations of the window's shattering in which the window doesn't shatter. So, on the revised account, neither will be appropriately cited as a cause, so long as there is some other event which does make a difference to whether the window shatters—as surely there is: witness the window's being fragile, the shutters being open, *etc*.

Perhaps the reductionist and the strong eliminativist could borrow a page from the weak eliminativist's playbook, and say that claims of the form 'C caused E' are either causal explanatory claims in disguise, or else pragmatically implicate that C causally explains E—what I called earlier the 'Davidson-Strevens thesis.' Unfortunately, that thesis does not pan out. In both its semantic and pragmatic flavors, it leads us

36 Similar points are made by Schaffer (2001) and Strevens (2003, 2008).

into bad predictions. Moreover, even if we accept the pragmatic flavor of the thesis, it does not help to explain the infelicity of 11.

In the first place, *C*'s being a part of an adequate explanation of *E* is not necessary for the truth (or the felicity—depending upon which version of the Davidson-Strevens thesis we are arguing against) of the causal claim '*C* caused *E*.' Suppose that you come to me with a bad case of insomnia and ask for something to help you rest better. I hand you a herb and tell you to eat it. You do, and shortly thereafter become sleepy. Suppose that it is common knowledge between us, in this context, that the herb brought about your sleepiness. If you then ask me 15,

15. Why did I get sleepy?

most philosophers of explanation, including Strevens, maintain that 16,

16. Because the herb has a dormitive virtue.

would be a poor answer. In this context, it is common knowledge between us that the herb brought about your sleepiness, and all that it is to have a dormitive virtue is to produce sleepiness. Therefore, in this context, the fact that the herb has a dormitive virtue provides no genuinely explanatory information about why you became sleepy. In this context, there is no adequate explanation of your sleepiness that makes reference to the herb's dormitive virtue. So 17

17. Your sleepiness is explained by the herb's having a dormitive virtue.

is false. Nevertheless, the causal claim 18

18. Your sleepiness was caused by the herb's having a dormitive virtue.

appears to be true, or at least felicitous, in this context. Even though citing the fact that the herb has a dormitive virtue doesn't adequately explain your sleepiness, the herb's having a dormitive virtue still *did cause* your sleepiness. If your intuitions vacillate here, consider 19:

19. Your sleepiness wasn't caused by the herb's having a dormitive virtue.

It would be perfectly appropriate, if somebody uttered 19, to correct them with an utterance of 18. So the adequacy of an explanation whose explanandum is *E* and whose explanans include *C* is not necessary for the truth, or the felicity, of a causal claim '*C* caused *E*'.

A causal reductionist or eliminativist might want to claim that the fact that the herb has a dormitive virtue does provide *some* explanatory information in this context, since to have a dormitive virtue, the herb must at least be such as to produce sleepiness in situations similar to this one. However, simply because the herb produced sleepiness in *this* case, it doesn't follow that the herb will produce sleepiness in other similar cases. Insofar as modal resiliency is an explanatory virtue, the fact that the herb has a dormitive virtue would provide at least some genuinely explanatory information in this case. In response to this objection, we may alter the case so that my response isn't 16, but rather 'Because the herb has a one-off dormitive virtue', where what it is to have a 'one-off dormitive virtue' is to be such as to produce sleepiness in *exactly these circumstances* (then, having a dormitive virtue entails having one-off dormitive virtue, though having a one-off dormitive virtue does not entail having a dormitive virtue). That the herb has a one-off dormitive virtue does not tell us anything about the herb's ability to produce sleepiness in similar situations or nearby possibilities. Then, the herb's having a one-off dormitive virtue does not provide any genuinely explanatory information about why you got sleepy. *Even so*, it seems correct to say, in this context, that your sleepiness was caused by the herb's having a one-off dormitive virtue. Had the herb not had a one-off dormitive virtue, you wouldn't have gotten sleepy.

Moreover, note that, even if the pragmatic version of Davidson-Strevens thesis were correct, it wouldn't be sufficient to explain the intuitive truth of 11. That's because, according to the pragmatic version of the Davidson-Strevens thesis, a causal claim 'C caused E' implicates that C explains E. However, negating a claim which implicates that p need not implicate that not-p. 'Some of the boys went to the lake' implicates that not all of the boys went to the lake. However, 'None of the boys went to the lake' certainly does not implicate that all of the boys went to the lake. So, even if 'The orbit of Gliese 163c caused stagflation' implicates that the orbit of Gliese 163c is explanatorily relevant to the economy's stagflating, this doesn't show that 'The orbit of Gliese 163c didn't cause stagflation' implicates that the orbit of Gliese 163c *isn't* explanatorily relevant to the economy's stagflating. Perhaps if it were obvious that Gliese 163c's orbit *did* cause stagflation, then we could tell some story about how a claim like 11 could be used to implicate that Gliese 163c's orbit does not *explain* stagflation—*e.g.*, uttering 11 flouts the maxim of quality. However, it is *far* from obvious that Gliese 163c's orbit caused stagflation (in fact, I'm fairly certain that it didn't; and I hope to have at least cast the reader into some doubt on this question). And given this, there is no reason to rule out that an utterance of 11 is intended to deny the existence of a *causal* relation—and not merely an *explanatory* relation—between the orbit of Gliese 163c and stagflation.

To sum up: Lewis's strategy for denying 11—we mistakenly judge 11 to be true because we usually neglect causes whose influence is negligible—founders because high-level events can have substantial and non-negligible effects upon the *way* in which another high-level event occurs, without intuitively causing that high-level event (as demonstrated by the infelicity of 13). The Davidson-Strevens strategy for denying 11—claims like 11 are either true in virtue of the fact that the cause does not explain the effect, or else they are false, but mistaken for true because they implicate the true claim that the cause does not explain the effect—founders because, firstly, causal explanatoriness is not necessary for either the apparent truth or the felicity of a causal claim, and secondly, in its pragmatic form, the thesis fails to explain the apparent truth of 11, since denying a claim which implicates that p need not implicate that not-p. More generally, given that it is far from obvious that there is a causal relation between Gliese 163c's orbit and stagflation, there's no reason for a listener to rule out that an utterance of 11 is meant to deny the existence of a causal relation between these two events.

2.4 In Summation

The straightforward strategies for explaining 11's apparent truth while maintaining its falsity all come up short. I conclude that appearances are not deceiving, and 11 actually is true. Since 12 is also true, causal reductionism and strong causal eliminativism are false. Additionally, we saw above that weak causal eliminativism is only as plausible as the Davidson-Strevens thesis, since it is this thesis which allows the weak eliminativist to avoid the absurd conclusions that 'carbon emissions caused climate change' is unassertible. Since the Davidson-Strevens thesis falters, so too does weak causal eliminativism. This leaves causal emergentism as the last position standing.

Of course, additional epicycles are always available; additional semantic or pragmatic theses could be introduced to pick up the slack left over by the Lewis and Davidson-Strevens theses. (As I said above, it is a criterion of bare adequacy on such an account that it provide some explanation of why it is that we would choose to talk in accord with such semantic or pragmatic theses.) It cannot be predicted in advance whether counterexamples to these further theses could be discovered. More effective than arguing against these epicycles, I think, is demonstrating that there is a more attractive alternative. And, as the example of Lewis's 1973 account demonstrates, in order to be causal emergentists, we need not deny that high-level causal facts reduce to, or obtain in virtue of, fundamental physical facts. We need only deny that they obtain exclusively in virtue of low-level *causal* facts. Such a metaphysical commitment is far from "extravagant", as Strevens claims. And theories of causation which entail causal emergentism are already well established and well regarded. Besides the account of Lewis (1973), by the way, causally emergentist theories of causation include the process theory of Dowe,[37] the regularity theory of Mackie,[38] the probabilistic theory of Suppes,[39] and the interventionist theory of Woodward.[40] (Or, at least, these theories generate causal emergentism is we are allowed to feed both high- and low-level events into their truth-conditions.) Causal emergentism allows us to account for the apparent truth of both 11 and 12 without the contortions of elaborate semantic or pragmatic theses. And it can do so without rustling any (microphysical) reductionist feathers. To my mind, this counts as a considerable point in its favor.

2.4.1 Looking Forward

In the following three chapters, I will advance an emergentist theory of singular causation. Before proceeding, I'd like to sketch, in broad outline, how this theory will end up entailing causal emergentism. On the theory to be developed here, causal relations presuppose the existence of what I will call *structural determination relations* between variables. Variables represent properties of the world; they stand to their values as determinables stand to determinates. For instance, *color* is a variable whose values are *red*, *blue*, *yellow*, *etc*. On my view, what it is for a structural determination

[37] Though I don't have the space to go into it here, the curious reader will find that Dowe more or less explicitly embraces causal emergentism in chapter 7 of his 2000.

[38] Simply because c is part of a minimally sufficient condition for e, this does not entail that C is a part of a minimally sufficient condition for E, as the case of Gliese 163c and stagflation readily illustrates.

[39] This is because it is possible for both $\Pr(e \mid c) \neq \Pr(e \mid \neg c)$ and $\Pr(E \mid C) = \Pr(E \mid \neg C)$ to be true, if we understand C and E to be events in the measurable space over which Pr is defined, and c and e to be events such that $c \subsetneq C$ and $e \subsetneq E$.

[40] Simply because there's an in-principle intervention on the maximally-fine-grained variable describing the fundamental physical state of Gliese 163c which brings about a change in the value of the maximally-fine-grained variable describing the fundamental physical state of stagflation, this does not mean that there is an in-principle intervention on a more coarse-grained variable describing Gliese 163c's orbit which brings about a change in the value of a coarse-grained variable describing whether or not the U.S. economy stagflates during the 1970's.

relation to obtain between two variables X and Y is, roughly, for there to exist a function, ϕ, such that the fundamental laws of nature map possibilities in which X takes on the value x to possibilities in which Y takes on the value $\phi(x)$, for every possible value x of X. But then note that the existence of a mapping from the values of a finely-grained variable X to the values of a finely-grained variable Y does not guarantee the existence of a mapping from the values of a more coarsely-grained variable U to the values of a more coarsely-grained variable V—even when the value of X entails the value of U and the value of Y entails the value of V. (See figure 2.) So, even when a variable U's taking on the value u reduces to a variable X's taking on the value x, and a variable V's taking on the value v reduces to a variable Y's taking on the value y, it doesn't follow from the fact that $X = x$ caused $Y = y$ that $U = u$ caused $V = v$.

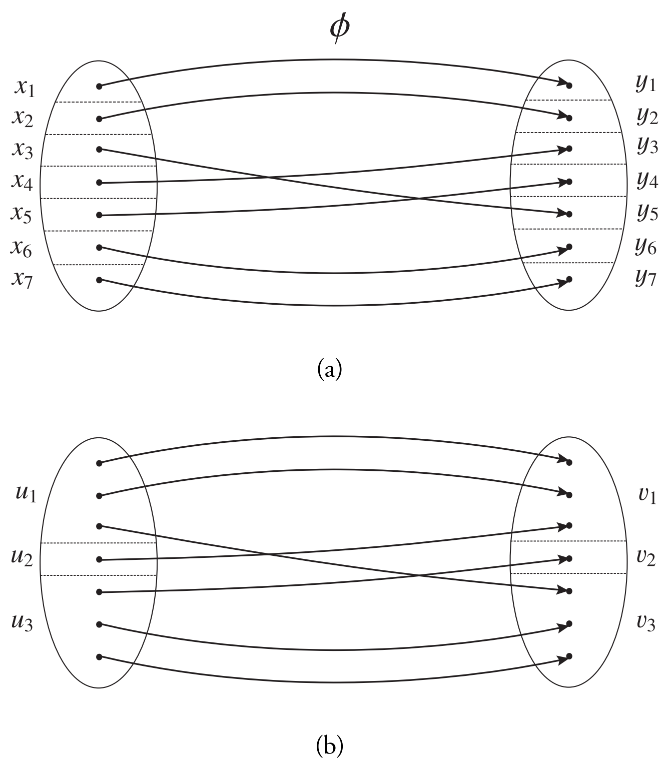


Figure 2: Just because there is a functional relationship ϕ between the fine-grained variables X and Y, 2(a), that does not mean that there is a functional relationship between the more coarse-grained variables U and V, 2(b).

CHAPTER 3

A THEORY OF STRUCTURAL DETERMINATION

3.1 Introduction

As a rough approximation, regularity theories of causation hold that, given the circumstances, causes are nomically *sufficient* for their effects. As a matter of law, if the cause is present in these circumstances, then the effect will be present too. As a rough approximation, counterfactual theories of causation hold that, given the circumstances, causes are nomically *necessary* for their effects. As a matter of law, were the cause to have been absent in these circumstances, so too would the effect have been absent. As Lewis (1973) notes, Hume appears to affirm both a regularity and a counterfactual account, saying first that a cause is "an object followed by another, and where all the objects, similar to the first, are followed by objects similar to the second", and then that a cause is an object followed by another, "where, if the first object had not been, the second never had existed."[1] In this chapter and in chapter 5, I will suggest that we follow Hume in accepting that causes are in some good sense

[1] Hume 1975, §VII.

both nomically sufficient and nomically necessary for their effects. On this dual theory, nomic sufficiency accounts are roughly correct about what I will call *structural determination*; while nomic necessity accounts are roughly correct about the relation of *singular causation*.

Singular causal relations, the relations expressed by sentences of the form *c's ϕ-ing caused e to ψ*, where *c*'s ϕ-ing and *e*'s ψ-ing are particular events or facts, are familiar philosophical fare. Structural determination, less so. As I will explain in further depth below, structural determination relations link localized qualities or quantities of different parts of the world.[2] We can represent these qualities or quantities with variable values. When we do so, structural determination relations are representable as structural *equations*, which establish functional relationships between the values of those variables.

Structural equations like these have become increasingly prevalent both in applied work in the special sciences and in the philosophical literature on causation. Within philosophy, these models have been used to provide novel semantics for causal counterfactual conditionals,[3] investigate traditional metaphysical questions about singular causation,[4] explicate the nature of causal enquiry in the special sciences,[5] and undergird novel statistical techniques for drawing inferences about the causal structure of the world on the basis of sample data.[6] However, relatively little has been said about how exactly to understand these structural equations models. That is, very little has been done to get clear about what exactly someone commits themselves to when they endorse one of these models—what exactly, that is, a structural equations model *says* about the world.[7] This chapter is my attempt to ameliorate that situation, to supply

---

[2] Contrasting structural determination with *singular* causation invites the interpretation that structural determination relations are general, or type, relations. To be clear: relations of structural determination relate *particular*, or token, qualities or quantities of the world.

[3] See, *e.g.*, Hiddleston (2005a), Shulz (2011), and Briggs (2012).

[4] See, *e.g.*, Hitchcock (2001), Woodward (2003, ch. 2), Menzies (2004), Halpern & Pearl (2001, 2005), Menzies (2007), and Weslake (ms)

[5] See, *e.g.*, Woodward (1999), Pearl (1999), Woodward & Hitchcock (2003a,b), and Woodward (2003, ch. 7)

[6] See, *e.g.*, Pearl (2000, 2009) and Spirtes et al. (2000)

[7] Some notable exceptions include Handfield et al. (2008), Baumgartner (2013), and Glynn (2013). Woodward (2003) provides a non-reductive account of structural equations models in terms of in-principle *interventions*, where what constitutes an intervention is cashed out in terms of structural equations models and singular causation; and singular causation is cashed out in terms of interventions and singular causation. While I do not think that Woodward (2003)'s account is viciously circular, and while it does tell us *something* about what somebody commits themselves to when they endorse a structural equations model, it tells us far less than we might have hoped. I also agree with Glynn

an understanding of the content of structural equations models.

To the extent that it has been discussed, the consensus view appears to be that structural equations represent patterns of causal counterfactual dependence among variable values.[8] (Not just any counterfactual is a *causal* counterfactual. While the question of which counterfactuals are causal is a question to be decided by theory rather than stipulation, at the least, causal counterfactuals must be *non-backtracking*,[9] and they must relate *distinct* events.[10].) In §3.3 below, I will explain why causal counterfactuals are not well-suited to provide a reduction basis for structural determination relations. My contention there will be that causal counterfactuals are not capable of securing the independent manipulability of the structural determination relations in a structural equations model—a property known as *modularity*.[11] In contrast, I will suggest that one variable is structurally determined by others just in case, within a certain region of modal space, the values of the latter variables are sufficient for the value of the former. Once this notion of structural determination has been specified, we will find that we can use structural determination relations to provide a semantics for causal counterfactuals. In chapter 5, I will use these causal counterfactuals to provide an account of singular causation. The upshot of the discussion there will be that Lewis's 1973 analysis of causation as the ancestral of counterfactual dependence was in large part correct. A few extra bells and whistles are necessary; however, once the counterfactuals have been properly grounded in the networks of structural determination codified in a correct structural equations model, the standard objections to Lewis's original counterfactual account will dissolve.

Here's how the rest of this chapter is going to go: In §3.2, I will introduce structural equations models. In §3.3, I will review and work to explicate the currently mainstream understanding of structural equations models, according to which they represent patterns of causal counterfactual dependence. In §3.3.1, I will demonstrate

(2013) that Woodward's reasons for forsaking a reductive account of structural equations models in terms of causal counterfactuals are not ultimately persuasive.

[8] In particular, see Hitchcock (2001), Hall (2007), Hitchcock (2007), Woodward & Hitchcock (2003a), Halpern & Hitchcock (2010), as well as the summary of Woodward & Hitchcock (2003a) provided in Menzies (2008), which fleshes out some of the claims of Woodward & Hitchcock in interesting ways. Glynn (2013) ably defends a reductive causal counterfactual account of structural equations models against Woodward (2003)'s arguments that such a reductive account is infeasible.

[9] See Lewis (1979b)

[10] See Kim (1973), Lewis (1986b)

[11] See Hausman & Woodward (1999) and Woodward (2003).

that this understanding of structural equations models is incapable of guaranteeing that structural determination relations are independently manipulable—that is, that the equations in a structural equations model are *modular*. This will lay the groundwork for §3.4, in which I present my preferred way of understanding structural equations models, which I dub the 'nomic sufficiency' account of structural determination. In §3.4.2, I will show how the nomic sufficiency account allows us to understand the technical notion of an *intervention* and how it guarantees the independent manipulability of the structural determination relations in a correct structural equations model. In §3.4.4, I will show how the semantics for causal counterfactuals provided by the nomic sufficiency account advances on the standard Lewis/Stalnaker semantics. In §3.4.5, I will consider and respond to a potential problem for the nomic sufficiency account. I'll conclude in §3.5.

3.2 Structural Equations Models

A structural equations model $\mathcal{M}$ is a triple $< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ of a vector of exogenous variable $\mathcal{U} = (U_1...U_m)$, a vector of endogenous variables $\mathcal{V} = (V_1...V_n)$ and a vector of structural equations $\mathcal{E} = (\phi_{V_1}...\phi_{V_n})$, one for each endogenous variable.[12] Formally, a variable is a partial function from a set of possibilities (or 'worlds') Ω to the real line $\mathbb{R}$. What makes the function *partial* is just that it needn't map each and every possibility $\omega \in \Omega$ to some real number. So, for instance, I might be interested in the variable S = *number appearing on the digital scale at* t. This variable assigns a value, s, to every world at which the scale displays a number at t. However, it will not assign any value to a world at which the scale does not display a number at t, or does not exist t. Here's another (equivalent) way to understand a variable: it is an assignment of values to a set of pairwise inconsistent propositions $\{\mathscr{P}_i\} \subset \wp(\Omega)$. Which value the variable takes on depends upon which of these propositions is true. For instance,

[12] A word on notation: throughout, I'll be using uppercase Latin letters $(A, B, C, ..., Z)$ to represent variables, and the corresponding lowercase letters $(a, b, c, ..., z)$ to stand for the values of those variables. Functions will be denoted with ϕ, with subscripts added to indicate which variable the function is associated with. Vectors will be denoted with boldface ($\mathbf{V}$, $\mathbf{v}$, *etc.*). At times (as in the sentence to which this note is appended) I will use the function name alone to denote the entire structural equation—for instance, I will write 'ϕ_Y' to denote the structural equation '$Y := \phi_Y(X_1, ..., X_N)$'. Propositions will be denoted with uppercase script letters $(\mathscr{A}, \mathscr{B}, \mathscr{C}, ...)$.

I'll also be slightly abusing set-theoretic notation, $\in, \cup, \subseteq, -$, and so on, by applying it to *vectors* of variables. So, '$\mathbf{V} \subseteq \mathbf{V}^*$' says that every variable in $\mathbf{V}$ is also in $\mathbf{V}^*$, '$\mathbf{V} \cup \mathbf{V}^*$' refers to a vector containing all the variables in $\mathbf{V}$ as well as all those in $\mathbf{V}^*$, and so on. (There will in general be many such vectors. It won't matter for my purposes which is denoted by these expressions. Pick one.)

the variable F could assign the value f to the proposition *you exert a force of f Newtons on the surface of the scale at t*, for every f in some specified range. If you exert a force of 5 Newtons on the surface of the scale at t, then $F = 5$; if you exert a force of 12 Newtons on the surface of the scale at t, then $F = 12$; and so on. In general, variables stand to their values as determinables stand to their determinants; just as being red is one way for an object to be colored, having the property represented by $V = v$ is one way for a part of the world to have the property represented by V. (A word on notation: I'll write '$V_\omega = v$' to mean that the value of ω, under the function V, is v. I will often use '$V = v$' to denote the proposition that V takes on the value v—*i.e.*, '$V = v$' denotes the set of worlds ω such that $V_\omega = v$.)

The structural equations in $\mathcal{E}$ establish functional relationships amongst the variables in $\mathcal{U} \cup \mathcal{V}$. For instance, suppose that the digital scale is accurate at t, zeroed out in the appropriate way, so that it reads '0' when subjected to the earth's gravitational force and the ambient air pressure, and nothing else (besides you) is exerting any force upon the surface of the scale. Then, the value of S will be *determined by* the value of F. If $F = 712$, then $S = 161$. If $F = 657$, then $S = 148$. In general, if $F = f$, then $S = \lceil 0.2248 \cdot f \rceil$.[13] (An object weighing 0.2248 pounds exerts 1 Newton on the surface of the Earth.) We can thus write down: $S := \phi_S(F)$, where

$$\phi_S(*) = \lceil 0.2248 \cdot * \rceil$$

'$S := \phi_S(F)$' is a *structural equation*. What makes it *structural* is that the equation is asymmetric; it matters which variable is to the left of the '$:=$'. That's because, in addition to claiming that the value of S is a function ϕ_S of the value of F, the structural equation makes the further claim that that the value of S is *determined by* the value of F in a way that the value of F isn't determined by the value of S. Which number is displayed on the scale at t is determined by the amount of force you exert on the surface of the scale at t. However, the amount of force you exert on the surface of the scale at t is not determined by which number is displayed on the scale at t. Here's a way of getting at this asymmetry: if there were a certain kind of intervention on the value of F—if, for instance, you were to put one foot on the floor—then the value of S would have been different—the scale would have displayed a different number. If, however, there were an intervention on the value of S—if, for instance, the scale was re-zeroed to read '0'—then the value of F would *not* be different—you would not

[13] '$\lceil x \rceil$' is the function which rounds x up to the closest integer.

$$F \longrightarrow S$$

Figure 3

suddenly exert 0 Newtons on the surface of the scale. (To emphasize this asymmetry, I use '$:=$' to distinguish that asymmetrical relation from the symmetrical '$=$'.)[14]

These variables and this structural equation together constitute a *structural equations model*, or a *causal model* (I'll use these terms interchangeably throughout). A *causal graph* provides a particularly intuitive and succinct way of representing much of the information contained in a causal model. A causal graph displays all of the determination relationships between the variables in $\mathcal{U} \cup \mathcal{V}$; if the value of a variable U determines the value of another variable V, then there will be an arrow with its tail at U and its head at V in the causal graph. For instance, the model of the scale's display and the force you exert on the scale's surface generates the causal graph shown in figure 3. This causal graph tells us that the value of F determines the value of S, without telling us exactly *how*. It tells us that the numbers on the scale's display are determined by the force you exert upon it, but it doesn't tell us what number will be displayed if you exert 500 Newtons on the scale; it doesn't, for instance, tell us what units the scale's display is set to. For that information, we'll have to look to S's structural equation in $\mathcal{E}$.

A causal model can involve many more variables and structural equations than this. Also, a single structural equation can relate more than two variables. Adapting an example from Pearl (2000, ch. 7), suppose that there are two riflemen, one standing on the left, the other standing on the right, who have their rifles aimed at a deserter. If the captain gives the order, then both riflemen will fire, and the deserter will die. We can model the causal structure of this case with $\mathcal{M}_4 =< (C), (L, R, D), \mathcal{E}_4 >$, where C is a binary variable which takes the value 1 if the captain gives the order to fire and takes the value 0 otherwise, L is a binary variable which takes the value 1 iff the left rifleman fires, R is a binary variable which takes the value 1 iff the right rifleman fires, and D is a binary variable which takes the value 1 iff the deserter dies. The structural

[14] It's worth noting that the functions ϕ_V must be *non-constant*. A constant function from one variable to another does not represent any kind of *determination of* of the latter variable by the former. (This restriction is important; leaving it out could make a difference to an account of singular causation; for instance, VanderWeele (2009) presents a counterexample to the Halpern & Pearl (2005) account of singular causation which turns on the fact that he models the system with constant functions.) We should also require that the domain of each structural equation include the entire image of their parent variables' structural equations, and *only* that image.

$$\mathcal{E}_4 = \begin{pmatrix} L := C \\ R := C \\ D := L \vee R \end{pmatrix}$$

$$\begin{tikzcd} L \arrow[r] & D \\ C \arrow[u] \arrow[r] & R \arrow[u] \end{tikzcd}$$

Figure 4

equations in $\mathcal{E}_4$ are shown in figure 4. (There, $\circ \vee *$ is the truth function $\max\{\circ, *\}$.) These equations tell us that the left rifleman will fire iff the captain gives the order, and likewise for the right rifleman. And the deserter will die iff at least one of the riflemen fire. $\mathcal{M}_4$ tells us that the value of C determines the values of L and R and that the values of L and R jointly determine the value of D .

When discussing a single structural equation $Y := \phi_Y(X_1, X_2, ..., X_n)$, it is common to refer to the left-hand-side variable Y as the *dependent* variable, and it is common to refer to the right-hand-side variables $X_1, X_2, ..., X_n$ as the *independent* variables. In the context of a structural equations model, however, $X_1, X_2, ..., X_n$ may each appear on the left-hand-side of their own structural equations. One structural equation's dependent variable may be another's independent variable. When we are talking about *systems* of structural equations, it is common to use the metaphor of genealogy to talk about the structural relationships between variables. Thus, all of the variables which *directly* determine the value of a given variable, V, (those which appear on the right-hand-side of V's structural equation, ϕ_V) are called V's *structural parents*. I'll use '$\mathbf{PA}(V)$' to refer to a vector of V's structural parents.[15] (If U is exogenous, then $\mathbf{PA}(U)$ is the empty vector.) In the model shown in figure 4, *e.g.*, $\mathbf{PA}(D) = (L, R)$. In a similar fashion, we can define V's *structural descendants*—with the slight wrinkle that we stipulate that every variable V is one of its own descendants. I'll use '$\mathbf{DE}(V)$' to refer to a vector of V's causal descendants. In the model shown in figure 4, $\mathbf{DE}(L) = (L, D)$.

A few paragraphs back, I invoked the notion of an *intervention*. Formally, an intervention is a way of setting the values of some of the variables in $\mathcal{U} \cup \mathcal{V}$ without directly affecting any of the other variables in $\mathcal{U} \cup \mathcal{V}$, or their determination structure. To illustrate, suppose that in the model shown in figure 4, the value of L is set to 1 via an intervention. Suppose, that is, that we perform an intervention to make the left rifleman fire—perhaps we bribe him. The way this is modeled is by replacing

[15] In general, for a variable V, there will be many vectors of V's structural parents, but it won't matter which one '$\mathbf{PA}(V)$' refers to. Pick one.

$$\mathcal{E}_{4,\,L=1} = \begin{pmatrix} L = 1 \\ R := C \\ D := L \vee R \end{pmatrix}$$

$$\begin{tikzcd} L \arrow[r] & D \\ C \arrow[r] & R \arrow[u] \end{tikzcd}$$

Figure 5

L's structural equation, $L := C$, with $L = 1$ (indicating that L has been set to 1 via an intervention) and leaving all other structural equations unchanged. We thus get the mutilated model[16] $\mathcal{M}_{4,\,L=1} =< (C), (L, R, D), \mathcal{E}_{4,\,L=1} >$, shown in figure 5. In $\mathcal{M}_{4,\,L=1}$, the value of L is no longer determined by the value of C. Whether the left rifleman fires is no longer determined by whether the captain gives the order. However, the value of D is still determined by the value of L. Whether the deserter dies is still determined by whether the left rifleman fires. In general, the graphical result of an intervention on a variable V is to remove all of the arrows leading *into* V (if such there be), to destroy all of the structural determination relations between V and $\mathbf{PA}(V)$, while leaving all other structural determination relations intact.

This property of a structural equations model—that there are in-principle hypothetical interventions upon the variables which leave all the other structural determination relations intact—is known as *modularity*.[17] Without modularity, structural equations models do not tell us anything about the results of hypothetical interventions, since without the assumption that the structural equations other than ϕ_V remain in place post-intervention, we cannot calculate the down-stream effects of setting the value of V.

Notice that not every way of setting the value of V will have this result. Some ways of setting the value of V will affect other variables in the graph as well. For instance, one way of setting S to 0, one way of making the scale read '0', is to simply lift you off of the scale. But this wouldn't count as an *intervention* on the value of S, since it wouldn't alter the manner in which the value of F determines the value of S. It wouldn't be correct to model this way of setting S to 0 by replacing $S := \phi_S(F)$ with $S = 0$, since the determination relation represented by $S := \phi_S(F)$ would still be in force. It would be this very determination relation that we would exploit in order to affect the value of S. Additionally, we could set the value of F in such a way that we affect the manner in which the value of F determines the value of S—*i.e.*, our

[16] The outstanding phrase 'mutilated model' comes from Pearl (2009).
[17] See Hausman & Woodward (1999) and Woodward (2003).

meddling could have the result of *changing* the structural equation ϕ_S. For instance, we might decide to keep you from stepping onto the scale by placing a dead five-pound rat on the scale. In that case, our method for setting the value of F would alter the structural equation ϕ_S, replacing it with $S := \phi'_S(F)$, where

$$\phi'_S(*) = \lceil 0.2248 \cdot * + 5 \rceil$$

So only certain methods of setting the value of a variable in a causal model will count as *interventions* on the value of that variable, in our technical sense.[18]

Once we have this method for modeling interventions, a method for evaluating causal counterfactual conditionals comes along for free. On this account, the counterfactual $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true at a world ω according to the model $\mathcal{M}$ just in case $\mathcal{M}_{\mathscr{A}}, \mathcal{U}_\omega \models \mathscr{C}$. That is: the counterfactual $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true at a world ω, according to the model $\mathcal{M}$, given the variable assignment $\mathcal{U} = \mathcal{U}_\omega$, iff $\mathscr{C}$ is true in the model that we get by mutilating $\mathcal{M}$, with the variable assignment $\mathcal{U} = \mathcal{U}_\omega$, so as to make $\mathscr{A}$ true. To illustrate: suppose that, in the causal model shown in figure 4, the actual value of C is 0. Suppose, that is, the the captain doesn't actually give the order to fire. Then, neither the left nor the right rifleman fires, and the deserter does not die. And suppose that we want to evaluate the causal counterfactual 'If the left rifleman were to have fired, then the deserter would have died'—or '$L = 1 \mathrel{\Box\!\!\rightarrow} D = 1$'. To evaluate this causal counterfactual, we simply perform an intervention on the value of L so as to make the antecedent true; we mutilate the model, so that the value of L is no longer determined by the value of C, we set L to 1, and then we calculate the values of R and D in the mutilated model in accordance with their structural equations. If the consequent comes out true in the mutilated model $\mathcal{M}_{4, L=1}$, then the counterfactual was true in the original model $\mathcal{M}_4$. According to $\mathcal{M}_4$, then, '$L = 1 \mathrel{\Box\!\!\rightarrow} D = 1$' is true. If the left rifleman were to fire, then the deserter would have died. Note that, without modularity, we would not be able to evaluate these counterfactual conditionals, since, without modularity, there is no guarantee that the downstream structural determination relations would remain intact post-intervention.

[18] That the structural equations need not be invariant under all methods of setting the values of the variables of interest is, in my opinion, one of the central lessons of the so-called *Lucas Critique* of macroeconometric policy evaluation—see Lucas (1976). For an excellent essay on the topic of how certain manipulations can change the underlying structural determination relations, and why such manipulations can nevertheless constitute effective strategies for achieving our ends, see Cartwright (2009).

As an aside: this framework additionally allows us to distinguish these causal counterfactuals from *backtracking* causal counterfactuals like 'if the left rifleman were to have fired, then the captain would have to have given the order to fire', and 'if the left rifleman were to have fired, then the right rifleman would have to have fired as well (since, after all, had the left rifleman fired, the captain would have to have given the order)'. While these counterfactuals have a true reading, they should also be distinguished from the causal counterfactuals above. In the framework of causal models, we can mark the distinction between these two kinds of counterfactuals as follows: while a causal counterfactual is evaluated by performing an intervention on the model so as to make the antecedent true and then solving for the other variables in the model, a *backtracking* causal counterfactual is evaluated by keeping the original model intact, making the antecedent true, and solving for the other variables in the model (if there is such a solution).[19]

An account of structural determination must explain why structural equations models have the properties they do. In particular, it must explain why they allow us to correctly evaluate causal counterfactual conditionals in this way, and it must explain why a correct system of structural equations is modular. It would be a benefit of an account if it could explain why only certain ways of setting the values of variables leave the downstream structural determination relations unaffected, as well as providing a principled way of distinguishing the ways of setting the values of variables which do from those which do not constitute *interventions*, in our technical sense. I will attempt to provide an account which is capable of meeting each of these explanatory demands. I will not attempt to account for probabilistic structural determination relations (of the sort that I believe are implicated in probabilistic causation). Nor will I be concerning myself with backwards structural determination relations, in which the future state of the world structurally determines the past state of the world. That's not because I think that there aren't, or couldn't be, probabilistic or backwards structural determination relations. Considering these issues here would simply muddy already murky waters. Another task for another day.

[19] In the event that there is no solution consistent with the truth of the antecedent, we could choose to say that the backtracking counterfactual is either false, truth-valueless, or trivially true. For instance, in model $\mathcal{M}_4$, there is no solution to the system of equations $\mathcal{E}_4$ within which $L = 1$ and $R = 0$. So there would be something defective about the backtracking counterfactual 'had the left rifleman fired and the right rifleman not fired, then the captain would have both given and not given the order to fire.' Perhaps considerations of theoretical economy would lead us to count it as trivially true, but there would be something defective about it all the same.

3.3 The Causal Counterfactual Understanding

Let's say that a structural equation $V := \phi_V(\mathbf{PA}(V))$ is *descriptively adequate* at world ω just in case $V_\omega = \phi_V(\mathbf{PA}(V)_\omega)$. Just as mere descriptive adequacy is not sufficient for a universal generalization to be a law of nature, mere descriptive adequacy is not enough for a structural equation to be correct. There must additionally be some kind of genuine *determination* of V by $\mathbf{PA}(V)$. So an account of structural equations models must say something about what it takes, beyond mere descriptive adequacy, for a structural equation to be correct.

One of the more popular ways of understanding structural equations models appeals to causal counterfactual conditionals.[20] Hitchcock articulates this view in his 2001:

> A system of structural equations is an elegant means for representing a whole family of counterfactuals...The correctness of a set of structural equations, and of the corresponding graph, depends upon the truth of these counterfactuals. (Hitchcock, 2001, p. 283–84).

On this account, what it is for an isolated structural equation $V := \phi_V(\mathbf{PA}(V))$ to be correct is just for it to be the case that, for any subvector $\mathbf{PA}^* \subseteq \mathbf{PA}(V)$, were $\mathbf{PA}^*$ to take on the values $\mathbf{pa}^*$, V would take on the value $\phi_V(\mathbf{pa}^*(V))$,

$$(V1) \qquad \forall \mathbf{PA}^* \subseteq \mathbf{PA}(V) \quad \forall \mathbf{pa}^* \;\; (\mathbf{PA}^* = \mathbf{pa}^* \;\; \Box\!\!\rightarrow \;\; V = \phi_V(\mathbf{pa}^*(V)))$$

where $\mathbf{pa}^*(V)$ is the assignment given to V's causal parents by $\mathbf{pa}^*$ (if $\mathbf{pa}^*$ doesn't assign any value to one of V's causal parents, then $\mathbf{pa}^*(V)$ gives that parent its actual value).

More generally, we can say that what it is for a structural equation ϕ_V, in the context of a causal model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$, to be correct at a world ω, is just for it to be the case that, for every subvector $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \stackrel{\text{def}}{=} \mathcal{U} \cup (\mathcal{V} - V)$, and any assignment of values $\overline{\mathbf{v}}^*$ to $\overline{\mathbf{V}}^*$, were $\overline{\mathbf{V}}^*$ to take on those values, V would take on the value $\phi_V(\mathbf{pa}_{\overline{\mathbf{v}}^*}(V))$

$$(\mathcal{V}1) \qquad \forall V \in \mathcal{V} \quad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \quad (\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \;\; \Box\!\!\rightarrow \;\; V = \phi_V(\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)))$$

where $\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$ assigns V's parents the values determined by $(\mathcal{U} - \overline{\mathbf{V}}^*)_\omega, \overline{\mathbf{v}}^*$, and $\mathcal{E} - \bigcup_i \{\phi_{U_i}\}$, for each endogenous $U_i \in \overline{\mathbf{V}}^*$. That is, $\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$ assigns $\mathbf{PA}(V)$ the values

[20] In addition to Hitchcock (2001), see Hall (2007), Hitchcock (2007), Woodward & Hitchcock (2003a), Menzies (2004, 2008), Halpern & Hitchcock (2010), and Glynn (2013).

determined by the mutilated model $\mathcal{M}_{\overline{\mathbf{V}}^*=\overline{\mathbf{v}}^*}$, with the actual assignment of values to the exogenous variables.[21]

The causal counterfactuals in ($V1$) and ($\mathcal{V}1$) are to be evaluated in the standard way (see Stalnaker 1968, Lewis 1973b). To evaluate a counterfactual $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$, we consider some privileged set of possibilities determined by $\mathscr{A}$ and the world of evaluation ω, and check to see whether $\mathscr{C}$ is true in those possibilities. Exactly which possibilities we ought to check is a complicated and controversial matter. However, for the most part, we can sidestep these issues here. We need only endorse the following general framework: there is a *selection function*, f, which is a function from pairs of propositions, $\mathscr{A}$, and worlds, ω, to sets of worlds, $f(\mathscr{A}, \omega)$. Whenever a counterfactual conditional $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true at a world ω, what makes it the case that $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true is that $f(\mathscr{A}, \omega) \models \mathscr{C}$.[22] Different accounts of the selection function will yield different truth conditions for counterfactual conditionals. Stalnaker interprets $f(\mathscr{A}, \omega)$ as the singleton containing the $\mathscr{A}$ world which is *most similar* to ω, on some suitable similarity metric. I personally favor the account of counterfactuals given by Maudlin (2007), according to which $f(\mathscr{A}, \omega)$ is the set of worlds generated by performing localized, surgical alterations to the state of ω so as to make $\mathscr{A}$ true at the appropriate time, and then time-evolving the state of those worlds into the future according to the fundamental laws of nature.[23] However, for my purposes, it won't matter what we say about f, so long as we agree that it satisfies the following

[21] Here and throughout, I'm using 'actual' as an indexical like 'here', and *not* as a rigid designator for the actual world.

[22] From the standpoint of Lewis (1973b)'s account of counterfactuals, it will appear that, by adopting this general framework, I am tolerating the so-called *limit assumption*—the assumption that, for any arbitrary antecedent $\mathscr{A}$ and world ω, there is a set of *most similar* $\mathscr{A}$-worlds from ω, that there is not an infinite sequence of ever-more-similar worlds (see Lewis 1973b, Stalnaker 1980). Appearances are deceiving. The limit assumption is not needed for any of my arguments here. In Lewis's framework, for any case in which the limit assumption fails and $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true, we can just define $f(\mathscr{A}, \omega)$ to be the largest sphere centered on ω containing at least one $\mathscr{A}$ world and throughout which the matieral conditional $\mathscr{A} \supset \mathscr{C}$ is true. So long as $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true, there will be some such sphere. If it is false, of course, there won't be such a sphere, so this won't do as an account of the truth conditions of these counterfactuals. However, I am not interested in providing truth conditions for these counterfactuals. Rather, I am interested in the question of whether the truth of a set of counterfactuals is sufficient to guarantee the correctness of a structural equations model. And this trick will tell us what we can infer from the truth of $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ at a world ω, on Lewis's account. From the truth of $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ at ω, we can infer that there is a sphere centered on ω containing an $\mathscr{A}$ world, and throughout which the material conditional $\mathscr{A} \supset \mathscr{C}$ is true.

[23] See Hall (2007) and Paul & Hall (2013) for a discussion of how Maudlin's semantics for counterfactuals can be used to provide truth conditions for structural equations models.

three properties.[24]

($f1$) $f(\mathscr{A}, \omega) \models \mathscr{A}$

($f2$) if $\mathscr{A} \models \mathscr{B}$ then $f(\mathscr{B}, \omega) \cap \mathscr{A} \subseteq f(\mathscr{A}, \omega)$

($f3$) at ω, there is a hypothetical *intervention* to set any $\mathbf{V} \subseteq \mathcal{U} \cup \mathcal{V}$ to $\mathbf{v}$ which yields a world $\omega_{\mathbf{V}=\mathbf{v}} \in f(\mathbf{V} = \mathbf{v}, \omega)$

($f1$) guarantees that the set of worlds at which we check to see what follows from the counterfactual supposition that $\mathscr{A}$ are all worlds at which this counterfactual supposition obtains. In terms of similarity, it guarantees that the most similar $\mathscr{A}$ worlds are all worlds at which $\mathscr{A}$ is true. And ($f2$) tells us that, if all $\mathscr{A}$ worlds are $\mathscr{B}$ worlds, then, if there are any $\mathscr{A}$ worlds among those that we consider when entertaining the counterfactual supposition that $\mathscr{B}$, then those worlds must also be considered when we entertain the counterfactual supposition that $\mathscr{A}$. In terms of similarity: if all $\mathscr{A}$ worlds are $\mathscr{B}$ worlds, then, if the most similar $\mathscr{B}$ worlds contain some $\mathscr{A}$ worlds, then those $\mathscr{A}$ worlds must be amongst the most similar $\mathscr{A}$ worlds.[25] ($f3$) connects the logic of counterfactuals to the notion of an *intervention*. It says that there is a hypothetical intervention to set some vector of variables $\mathbf{V}$ to the values $\mathbf{v}$ which yields one of the $\mathbf{V} = \mathbf{v}$ worlds we consider when entertaining the counterfactual supposition that $\mathbf{V} = \mathbf{v}$. ($f3$) may seem a bit odd if we think about $f(\mathscr{A}, \omega)$ as the set of $\mathscr{A}$ worlds most similar to ω. Going back to the causal model shown in figure 4, suppose, for the sake of illustration, that there are only two ways of setting the value of L to 1. We could either bribe the left rifleman to fire or we could bribe the captain to give the order. Only the former counts as an *intervention* setting the value of L to 1. But why shouldn't we think that the worlds where we bribe the left rifleman are less similar than the worlds where we bribe the captain? Perhaps the left rifleman would take a greater risk by firing, and would therefore require a higher bribe than the captain. In that case, it appears as though the worlds in which we bribe the left rifleman are less similar than the worlds in which we bribe the captain.[26] In any event,

[24] A dialectical note: in evaluating the causal counterfactual understanding of structural determination, I will be happy to grant the causal counterfactual theorist more than ($f1$)–($f3$). In laying down these conditions, I am not, therefore, tying the counterfactual theorist's hands behind their back before the fight. However, when I appropriate the counterfactual theorist's selection function for my own ends in §3.4, I will only require ($f1$)–($f3$).

[25] ($f1$) corresponds to Lewis's 2nd condition on the selection function; and ($f2$) is a weakening of Lewis's 4th condition. See Lewis 1973b, p. 58.

[26] To re-frame the objection in terms of Lewis (1979b)'s semantics: perhaps the miracles required

$$\mathcal{E}_6 = \begin{pmatrix} L := C \\ D := L \end{pmatrix}$$

$$C \longrightarrow L \longrightarrow D$$

Figure 6

the causal counterfactual account needs ($f3$) to be true. Without ($f3$), that account would not be able to use the correctness of a system of structural equations to infer anything about the results of hypothetical interventions.

Returning to ($\mathcal{V}1$): that condition imposes two constraints on a structural equation, ϕ_V, in the context of a causal model $\mathcal{M}$. In the first place, it says that the value of ϕ_V's left-hand-side variable, V, is sensitive to the values of ϕ_V's right-hand-side variables, $\mathbf{PA}(V)$, and they are sensitive in precisely the way specified by ϕ_V. Changes in the values of those variables would lead to changes in the value of V, and they would lead to precisely the changes specified by ϕ_V. Additionally, it says that the value of V is *only* directly sensitive to the values of $\mathbf{PA}(V)$. Holding fixed those values, changes in the values of the other variables in the model would not lead to changes in the value of V.

Note that ($\mathcal{V}1$) is stronger than the conjunction of each of the quantified counterfactuals ($V1$), for each endogenous variable $V \in \mathcal{V}$. ($\mathcal{V}1$) guarantees that hypothetical interventions on the value of a variable, X, in a causal model would determine the values of $\mathbf{DE}(X)$ in accordance with the equations in $\mathcal{E} - \{\phi_X\}$. In contrast, the conjunction of each of the quantified counterfactuals ($V1$) associated with each $V \in \mathcal{V}$ does not guarantee that hypothetical interventions on the value of X will determine the values of $\mathbf{DE}(X)$ in accordance with the equations in $\mathcal{E} - \{\phi_X\}$. To illustrate: suppose that the right rifleman takes the day off, so that the causal model shown in figure 6 correctly describes the structural determination relations between the captain's giving the order (C), the left rifleman's firing (L), and the deserter's dying (D). Suppose that, at the actual world, the captain doesn't give the order. Given the method for evaluating causal counterfactuals introduced in the previous section, this model entails that

(3.1) $$C = 1 \;\; \Box\!\!\rightarrow \;\; D = 1$$

If the captain had given the order, then the deserter would have died.

However, (3.1) does not follow from the truth of ($V1$), for each of the isolated

to get the captain to give the order are much smaller and more localized than the miracles required to get the left rifleman to give the order.

structural equations in $\mathcal{E}_6$,

$$C = 0 \mathrel{\Box\!\!\rightarrow} L = 0$$

$$C = 1 \mathrel{\Box\!\!\rightarrow} L = 1$$

$$L = 0 \mathrel{\Box\!\!\rightarrow} D = 0$$

$$L = 1 \mathrel{\Box\!\!\rightarrow} D = 1$$

since the counterfactual conditional is not transitive. Simply because, had the captain given the order, the left rifleman would have fired, and, had the left rifleman fired, the deserter would have died, this does not guarantee that, had the captain given the order, the deserter would have died. Perhaps the captain has a loud voice which distracts the rifleman and makes him miss his target. And perhaps the deserter is a good friend of the captain, so that the world where the left rifleman is bribed to shoot is much more similar to the actual world than the world in which the left rifleman shoots because the captain gave the order. In that case, it could be true that, had the captain given the order, then the left rifleman would have fired (though he would have missed, and the deserter would have lived). And it could also be true that, had the left rifleman shot (because he was bribed, not because the captain gave the order), the deserter would have died. Nevertheless, it would be false that, had the captain given the order, the deserter would have died. Had the captain given the order, the rifleman would have missed, and the deserter would have lived.

($\mathcal{V}1$), by contrast, will require, *inter alia*, that both (3.2) and (3.3) be true in order for the structural equations model to be correct.

$$C = 1 \mathrel{\Box\!\!\rightarrow} L = 1 \tag{3.2}$$

$$C = 1 \wedge L = 1 \mathrel{\Box\!\!\rightarrow} D = 1 \tag{3.3}$$

And (3.2) and (3.3) *do* entail (3.1), given ($f1$) and ($f2$).[27]

However, the truth of ($\mathcal{V}1$) is not sufficient for the correctness of a structural equations model. Take the familiar example of Suzy and Billy throwing their rocks at a window. Both Suzy and Billy have excellent aim, so if either of them throws their rock, then the window will shatter; and the window is sturdy enough so that

[27] Let $\mathscr{C} \overset{\text{def}}{=} C = 1$, $\mathscr{L} \overset{\text{def}}{=} L = 1$, $\mathscr{D} \overset{\text{def}}{=} D = 1$. Then, (3.2) is $\mathscr{C} \mathrel{\Box\!\!\rightarrow} \mathscr{L}$ and (3.3) is $\mathscr{C} \wedge \mathscr{L} \mathrel{\Box\!\!\rightarrow} \mathscr{D}$. ($f1$) tells us that $f(\mathscr{C}, \omega) \models \mathscr{C}$, and (3.2) tells us that $f(\mathscr{C}, \omega) \models \mathscr{L}$. Thus, $f(\mathscr{C}, \omega) \models \mathscr{C} \wedge \mathscr{L}$. Thus, $f(\mathscr{C}, \omega) \cap (\mathscr{C} \wedge \mathscr{L}) = f(\mathscr{C}, \omega)$. And since $\mathscr{C} \wedge \mathscr{L} \models \mathscr{C}$, ($f2$) tells us that $f(\mathscr{C}, \omega) \cap (\mathscr{C} \wedge \mathscr{L}) \subseteq f(\mathscr{C} \wedge \mathscr{L}, \omega)$; so $f(\mathscr{C}, \omega) \subseteq f(\mathscr{C} \wedge \mathscr{L}, \omega)$. And since $f(\mathscr{C} \wedge \mathscr{L}, \omega) \models \mathscr{D}$ (from (3.3)), $f(\mathscr{C}, \omega) \models \mathscr{D}$.

$$\mathcal{E}_7 = (W := S \vee B)$$

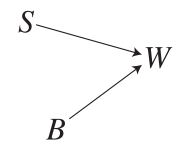


Figure 7

if neither of them throw their rock, then the window will not shatter. Suppose that Suzy actually throws and Billy doesn't, and that (3.4–3.11) are all true.

$$B = 0 \wedge S = 0 \mathrel{\Box\!\!\rightarrow} W = 0 \tag{3.4}$$

$$B = 0 \wedge S = 1 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.5}$$

$$B = 1 \wedge S = 0 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.6}$$

$$B = 1 \wedge S = 1 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.7}$$

$$S = 1 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.8}$$

$$S = 0 \mathrel{\Box\!\!\rightarrow} W = 0 \tag{3.9}$$

$$B = 1 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.10}$$

$$B = 0 \mathrel{\Box\!\!\rightarrow} W = 1 \tag{3.11}$$

(where B takes the value 1 if Billy throws and 0 otherwise, S takes the value 1 if Suzy throws and 0 otherwise, and W takes the value 1 if the window shatters and 0 otherwise). If the truth of ($\mathcal{V}1$) were sufficient for the correctness of a structural equations model, then the system of structural equations $\mathcal{E}_7$, shown in figure 7, would have to be correct.

However, this structural equations model says more than the counterfactuals (3.4–3.11) do. This models entails that were Billy to have thrown, Suzy (still) would have, $B = 1 \mathrel{\Box\!\!\rightarrow} S = 1$. But it is consistent with the truth of (3.4–3.11) that Suzy wouldn't have thrown her rock if Billy had thrown his. The problem here is that, even though the counterfactuals (3.4–3.11) are all true, there is *another* counterfactual that we must consider in order to know whether Suzy would have thrown if Billy hadn't, namely $B = 1 \mathrel{\Box\!\!\rightarrow} S = 0$. If this counterfactual is true, then the value of S is determined by the value of B. And if S is determined by B, then $\mathcal{M}_7$ would be incorrect in virtue of its missing a necessary determination relation.

We can fix this problem by requiring that each exogenous variables is counterfactually independent of all the other variables in the model—*i.e.*, at a world ω, for any exogenous variable $U \in \mathcal{U}$, and any assignment of values $\overline{\mathbf{u}}^*$ to any subvector

$\overline{\mathbf{U}}^* \subseteq \overline{\mathbf{U}} \stackrel{\text{def}}{=} (\mathcal{U} - U) \cup \mathcal{V}$, were $\overline{\mathbf{U}}^*$ to take on those values, U would (still) take on its actual value, U_ω.

$$(\mathcal{U}1) \qquad \forall U \in \mathcal{U} \quad \forall \overline{\mathbf{U}}^* \subseteq \overline{\mathbf{U}} \quad \forall \overline{\mathbf{u}}^* \quad (\overline{\mathbf{U}}^* = \overline{\mathbf{u}}^* \mathrel{\Box\!\!\rightarrow} U = U_\omega)$$

This gives us the following account of the correctness of a structural equations model $\mathcal{M}$:

($\mathcal{M}1$)

$\mathcal{M}$ is correct at ω iff:

1. $\mathcal{V}$ satisfies ($\mathcal{V}1$) at ω
2. $\mathcal{U}$ satisfies ($\mathcal{U}1$) at ω

That is: a model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ is correct iff i) the endogenous variables $V \in \mathcal{V}$ counterfactually depend upon the variables in $\mathbf{PA}(V)$ in precisely the manner specified by $\phi_V \in \mathcal{E}$, ii) holding fixed the value of $\mathbf{PA}(V)$, they *don't* counterfactually depend upon the variables in $(\mathcal{U} \cup \mathcal{V}) - \mathbf{PA}(V)$, and iii) the exogenous variables $U \in \mathcal{U}$ are counterfactually independent of the rest of the variables in the model.

3.3.1 Problems with Modularity

Problems remain. In order for a structural equations model to be correct, the equations in $\mathcal{E}$ must be *modular*—that is, that there be in-principle interventions to set the values of any subset of $\mathcal{V}$ which leaves the structural equations of the non-intervened-upon variables intact. The problem is that modularity does not follow from ($\mathcal{M}1$) alone; nor can we formulate the requirement of modularity in terms of any finite number of causal counterfactual conditionals.

Distinguish two kinds of modularity: weak and strong. According to weak modularity, when there is an intervention or interventions to set the values of variables, the structural equations of the non-intervened-upon endogenous variables in the model will still be descriptively adequate. That is, when we perform hypothetical interventions on the values of the variables in $\mathcal{U} \cup \mathcal{V}$, taking us to the world ω_I, then, for every non-intervened-upon endogenous variable V,

$$\omega_I \models V = \phi_V(\mathbf{PA}(V))$$

According to strong modularity, whenever there is an intervention or interventions to set the values of variables, the structural equations of the non-intervened-upon endogenous variables in the model will still be *correct*. That is, when we perform hypothetical interventions on the values of the variables in $\mathcal{U} \cup \mathcal{V}$, taking us to the world ω_I, for every non-intervened-upon endogenous variable V,

$$\omega_I \models V := \phi_V(\mathbf{PA}(V))$$

On the counterfactual account, this means that it must at least be true that, for every non-intervened-upon endogenous variable V,

$$(3.12) \qquad \omega_I \models \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \ \forall \overline{\mathbf{v}}^* \ \ (\ \overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \ \ \Box\!\!\rightarrow \ \ V = \phi_V(\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)))$$

($\mathcal{M}1$) does not guarantee strong modularity because it does not guarantee the truth of (3.12).

To see why, consider again the structural equations model shown in figure 6. For that model, ($\mathcal{M}1$) requires that $C = 1 \ \Box\!\!\rightarrow D = 1$: at all the closest worlds at which the captain gives the order, the deserter will die, $f(C = 1, @) \models D = 1$. However, ($\mathcal{M}1$) does not require that $f(C = 1, @) \models L = 0 \ \Box\!\!\rightarrow D = 0$. It does not require that, at all the closest worlds at which the captain gives the order, whether the deserter dies counterfactually depends upon whether the left rifleman fires. And that means that ($\mathcal{M}1$) does not require that, at the worlds at which a hypothetical intervention setting C to 1 occurs, D's value is still structurally determined by L's value. So ($\mathcal{M}1$) will not guarantee that the equations in $\mathcal{E}$ are strongly modular.

Weak modularity is not modularity enough. Structural equations don't merely represent accidentally true patterns amongst variable values. They represent *determination* relations between variable values. To say that the structural equation $V := \phi(\mathbf{PA}(V))$ is unaffected by an intervention on another variable should be to say that the *determination* of V by $\mathbf{PA}(V)$ is unaffected. While weak modularity will be enough to tell us what values the variables in $\mathbf{DE}(X)$ will take on if there were an intervention on X, it will not be enough to tell us about the truth of nested counterfactuals of the form $\mathscr{A} \ \Box\!\!\rightarrow (\mathscr{B} \ \Box\!\!\rightarrow \mathscr{C})$, since, at the world at which the hypothetical intervention making $\mathscr{A}$ true occurs, there's no guarantee that $\mathscr{B} \ \Box\!\!\rightarrow \mathscr{C}$ will be true.

Here's a thought about how to achieve strong modularity: we don't merely require that ($\mathcal{V}1$) be satisfied. We additionally require that $\mathcal{V}$ meets the following condition.

($\mathcal{V}2$)

$$\forall V \in \mathcal{V} \quad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \Big[\ \overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \ \Box\!\!\rightarrow$$
$$\Big[\forall W \in \mathcal{V} - \overline{\mathbf{V}}^* \quad \forall \overline{\mathbf{W}}^* \subseteq \overline{\mathbf{W}} \quad \forall \overline{\mathbf{w}}^* \ (\overline{\mathbf{W}}^* = \overline{\mathbf{w}}^* \ \Box\!\!\rightarrow W = \phi_W(\mathbf{pa}_{\overline{\mathbf{w}}^*}(W)))\Big]\Big]$$

($\mathcal{V}2$) says that, if there were an intervention to set the values of any set of variables $\overline{\mathbf{V}}^*$, then ($\mathcal{V}1$) would still hold for all the non-intervened upon variables. This solution is not satisfying. With this new account, we are told that what it is for a structural equations model to be correct is, *inter alia*, for *both* ($\mathcal{V}1$) *and* ($\mathcal{V}2$) to be satisfied. But while ($\mathcal{V}2$) guarantees that, at the world where the hypothetical intervention occurs, ($\mathcal{V}1$) will hold, we have as yet no guarantee that, at that world, ($\mathcal{V}2$) will be satisfied. But if what it is for a structural equation to be correct is for both ($\mathcal{V}1$) and ($\mathcal{V}2$) to hold, then this account fails to guarantee that the structural equation will still be correct post-intervention; that is, it fails to secure strong modularity.

It's actually a bit worse than that. ($\mathcal{M}1$) cannot even guarantee *weak* modularity. To begin to see why, consider the following inference,

(EXPORTATION) $$\frac{(\mathscr{A} \wedge \mathscr{B}) \mathrel{\Box\!\!\rightarrow} \mathscr{C}}{\mathscr{A} \mathrel{\Box\!\!\rightarrow} (\mathscr{B} \mathrel{\Box\!\!\rightarrow} \mathscr{C})}$$

in the special case where $\mathscr{A}$ and $\mathscr{B}$ are compossible conjunctions of variable values, EXPORTATION is valid on the semantics given by the structural equations models.[28] $(\mathscr{A} \wedge \mathscr{B}) \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is evaluated by mutilating the model so as to set the variable values $\mathscr{A} \wedge \mathscr{B}$, and then checking to see whether $\mathscr{C}$ is true. $\mathscr{A} \mathrel{\Box\!\!\rightarrow} (\mathscr{B} \mathrel{\Box\!\!\rightarrow} \mathscr{C})$ is evaluated by mutilating the model so as to set the variable values $\mathscr{A}$, and then mutilating the model so as to set the variable values $\mathscr{B}$, and then checking to see whether $\mathscr{C}$ is true. So long as $\mathscr{A}$ and $\mathscr{B}$ are compossible conjunctions of variable values, the model that one gets by *first* setting the values $\mathscr{A}$ and *then* setting the values $\mathscr{B}$ is just the model that one gets by setting the values $\mathscr{A} \wedge \mathscr{B}$. So EXPORTATION will be valid in this special case, on the semantics provided by the structural equations models.

However, EXPORTATION is not valid on the standard semantics, even when $\mathscr{A}$ and $\mathscr{B}$ are unary and compossible.[29] In STALNAKER's terms, the closest $\mathscr{A} \wedge \mathscr{B}$ world need not be the closest $\mathscr{B}$ world from the closest $\mathscr{A}$ world. In other words, there's no

[28] On the semantics provided by BRIGGS (2012) (the only author I've encountered who extends the structural equation semantics for counterfactuals to include disjunctive antecedents), EXPORTATION will not be valid if $\mathscr{A}$ and $\mathscr{B}$ are allowed to be disjunctive. For example, in the system of equations $\mathcal{E} = (X := \phi_X(Z), Y := \phi_Y(Z))$,

$$((X = x \vee Y = y) \wedge (X = x \vee Y = y')) \mathrel{\Box\!\!\rightarrow} X = x$$

(where $y' \neq y$) will be true (the antecedent is logically equivalent to $X = x$), while

$$(X = x \vee Y = y) \mathrel{\Box\!\!\rightarrow} ((X = x \vee Y = y') \mathrel{\Box\!\!\rightarrow} X = x)$$

will be false.

[29] See MCGEE (1985) for a discussion of whether EXPORTATION is in fact valid for the counterfactual conditional.

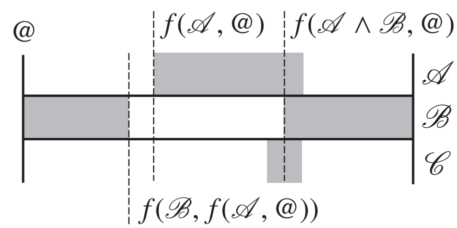


Figure 8: EXPORTATION is invalid on the STALNAKER semantics. In the diagram, modal space is spread out horizontally. Similarity is represented in terms of distance. The shaded area to the left of '$\mathscr{A}$' represents the possible worlds in which $\mathscr{A}$ is true. Similarly for the shaded areas to the left of '$\mathscr{B}$' and '$\mathscr{C}$.' In the diagram, then, the closest $\mathscr{A} \wedge \mathscr{B}$ world is a $\mathscr{C}$ world. However, the closest $\mathscr{B}$ world from the closest $\mathscr{A}$ world is not a $\mathscr{C}$ world.

guarantee that

$$f(\mathscr{A} \wedge \mathscr{B}, @) = f(\mathscr{B}, f(\mathscr{A}, @))$$[30]

And thus, it is possible that $f(\mathscr{A} \wedge \mathscr{B}, @) \models \mathscr{C}$, even while $f(\mathscr{B}, f(\mathscr{A}, @)) \not\models \mathscr{C}$. (An example is shown in figure 8.)

By the same token, it could easily turn out that, for any three variables V_1, V_2, and V_3,

$$f(V_1 = v_1 \wedge V_2 = v_2, @) \models V_3 = \phi_{V_3}(\mathbf{PA}(V_3)) \tag{3.13}$$

even though

$$f(V_2 = v_2, f(V_1 = v_1, @)) \not\models V_3 = \phi_{V_3}(\mathbf{PA}(V_3)) \tag{3.14}$$

While ($\mathcal{M}1$) guarantees (3.13), it is consistent with (3.14). But this means that ($\mathcal{M}1$) fails to guarantee that the equations in $\mathcal{E}$ will even be *descriptively adequate* after multiple sequential interventions. And the number of potential interventions is unbounded (we can always just set the value of X to x, then set it to $x' \neq x$, then set it back to x, then back to x', and so on and so forth, indefinitely). So there is no finite number of causal counterfactuals that is sufficient to guarantee that the equations in $\mathcal{E}$ are even weakly modular.[31]

[30] Above, I didn't define f for *sets* of worlds. Let's say that $f(\mathscr{B}, f(\mathscr{A}, @))$ is the union of $f(\mathscr{B}, \omega)$ for every $\omega \in f(\mathscr{A}, @)$.

[31] We might think that the fact that there is no *finite* number of causal counterfactuals that is sufficient to guarantee modularity isn't that worrying, since we could always just give a theory consisting of a counterfactual *schema* which entailed the requisite infinite number of counterfactuals. (Thanks to Jim Joyce for raising this objection.) In essence, this is just what the nomic sufficiency account does—

3.4 The Nomic Sufficiency Understanding

In this section, I will suggest that we can retain all of the virtues of the causal counterfactual understanding of structural equation models, without running into the problems with modularity raised in §3.3.1 above, by moving to an understanding of structural equations according to which what makes them correct is that they are descriptively adequate throughout an area of modal space meeting certain constraints. That is, an isolated structural equation $V := \phi_V(\mathbf{PA}(V))$ is correct just in case, for every world ω in some set of worlds $\mathfrak{F}_V$, $V_\omega = \phi_V(\mathbf{PA}(V)_\omega)$.

A useful orienting picture here is Mackie's notion of a *causal field*. In his 1965, Mackie states that causal claims must be evaluated relative to a set of alternate states of affairs within which the causes are parts of an occurrent minimally sufficient condition for the effect.[32] He calls this set of alternative states of affairs the *causal field*. This is roughly how I am thinking of the set of worlds $\mathfrak{F}_V$. Just as, on Mackie's account, the causes are the parts of an occurent minimally sufficient condition for the effect within the causal field, on the nomic sufficiency account, the values of a variable V's causal parents, $\mathbf{PA}(V)$, must be minimally sufficient for the value of V within $\mathfrak{F}_V$.[33] In virtue of this resemblance, I will call the set of possibilities $\mathfrak{F}_V$ V's *causal field*.

Of course, this is far too rough. For any structural equation ϕ_V, it will be easy to find *some* set of worlds $\mathfrak{F}_V$ within which ϕ_V is descriptively adequate. A structural equation according to which my height structurally determines the size of the earth will be descriptively adequate throughout $\mathfrak{F}_V$ if I only include worlds in $\mathfrak{F}_V$ in which the earth's diameter is a constant multiple of my height. But my height does not determine the size of the earth. The question of what possibilities to consider when evaluating the determination of one variable by another is a complicated one, but it is one that is faced by the nomic sufficiency account and the causal counterfactual

the recursive ($\mathfrak{F}2$) entails an infinite number of Stalnaker-style counterfactuals. So while I have a hard time seeing how to formulate a schematic counterfactual account which avoids the nomic sufficiency account's talk of causal fields, I suspect that the account would end up being very similar to the nomic sufficiency account. However, the resulting account would lack the benefit of the nomic sufficiency account which I discuss in §3.4.4.

[32] The condition is *minimally* sufficient for the effect just in case no subset of the condition is also sufficient for the effect. The minimal sufficient condition is *occurrent* iff it actually obtained on the occasion in question.

[33] The values of $\mathbf{PA}(V)$ are *sufficient*, and not (or not necessarily) necessary, for the value of V because two different assignments of values to $\mathbf{PA}(V)$ could get mapped by ϕ_V to the very same value of V. $\mathbf{PA}(V)$ must be *minimally* sufficient for V's value because we require that ϕ_V be a non-constant function of each of its parents. See fn 14.

account both. The causal counterfactual account solves it by appeal to some suitable selection function f. And I see no reason why the nomic sufficiency account cannot similarly avail itself of this very selection function—whichever one we fancied for the causal counterfactual understanding of structural equations models—to characterize the worlds which must be included in $\mathfrak{F}_V$.

As a first step, if we're considering an isolated structural equation $V := \phi_V(V)$ at a world ω, then we can require that, for every assignment $\mathbf{pa}^*$ to any $\mathbf{PA}^* \subseteq \mathbf{PA}(V)$, every world in $f(\mathbf{PA}^* = \mathbf{pa}^*, \omega)$ must be included in $\mathfrak{F}_V$. Similarly, if we're considering a structural equation $V := \phi_V(V)$ in a causal model $\mathcal{M}$ at a world ω, then we can require that, for every assignment of values $\overline{\mathbf{v}}^*$ to any $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, all the worlds in $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$ must be included in $\mathfrak{F}_V$.

$$(\mathfrak{F}1) \qquad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \quad f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \subseteq \mathfrak{F}_V$$

Putting this together with the requirement that ϕ_V be descriptively adequate throughout $\mathfrak{F}_V$, we can say that a causal model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ is correct at a world ω only if ($\mathcal{V}3$).

$$(\mathcal{V}3) \qquad \forall V \in \mathcal{V} \quad \exists \mathfrak{F}_V \text{ such that } \mathfrak{F}_V \text{ satisfies } (\mathfrak{F}1) \text{ and } \mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$$

If we stop here, then, in the presence of ($\mathcal{U}1$), we get an account which is equivalent to the causal counterfactual account's ($\mathcal{V}1$). That is, given ($\mathcal{U}1$), $\mathcal{M}$ satisfies ($\mathcal{V}1$) iff $\mathcal{M}$ satisfies ($\mathcal{V}3$). (**Theorem 1**, proved in §3.6, establishes the equivalence.)

Since this condition on the endogenous variables is equivalent to the causal counterfactual account's, if we stop here, we will run into the problems with modularity that we encountered in §3.3.1. However, we *needn't* stop here. We can additionally require that the condition imposed by ($\mathfrak{F}1$) holds, not only for the world of evaluation, but for every *other* world in $\mathfrak{F}_V$ as well.

$$(\mathfrak{F}2) \qquad \forall \omega \in \mathfrak{F}_V \quad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \quad f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \subseteq \mathfrak{F}_V$$

This amounts to the requirement that the set $\mathfrak{F}_V$ is closed under counterfactual suppositions about the values of any of the variables in $\overline{\mathbf{V}}$. At any world $\omega \in \mathfrak{F}_V$, making counterfactual suppositions about the values of any of the the variables in $\overline{\mathbf{V}}$ will deliver a set of worlds *inside* of $\mathfrak{F}_V$.

Putting this together with the requirement that ϕ_V be descriptively adequate throughout $\mathfrak{F}_V$, we get an account according to which a causal model $\mathcal{M}$ is correct at a world

ω only if ($\mathcal{V}4$).

($\mathcal{V}4$) $\forall V \in \mathcal{V} \quad \exists \mathfrak{F}_V \ni \omega$ such that $\mathfrak{F}_V$ satisfies ($\mathfrak{F}2$) and $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$

A structural equation belonging to a causal model satisfying ($\mathcal{V}4$) will continue to belong to a causal model satisfying ($\mathcal{V}4$) after any number of interventions to set the values of any of the other variables in the model. In §3.6, I offer a proof of the following theorem.

Theorem 2. *If ϕ_V belongs to a causal model satisfying ($\mathcal{V}4$) at a world ω_0, then ϕ_V will continue to belong to a causal model satisfying ($\mathcal{V}4$) after any number of consecutive hypothetical interventions to set the values of any $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \stackrel{\text{def}}{=} (\mathcal{U} \cup \mathcal{V}) - V$.*

This means that the nomic sufficiency account is not subject to the objection I raised for the counterfactual account in the previous section—*viz.*, that it could not guarantee that a structural equation ϕ_V would continue to be correct after multiple sequential interventions to set the values of the variables other than V.

3.4.1 Exogenous Independence

In §3.3, we saw that a structural equations model says more than just that the endogenous variables $V \in \mathcal{V}$ are structurally determined by their parents, and are not structurally determined by any of the other variables in $\mathcal{U} \cup \mathcal{V}$. It additionally says that the exogenous variables aren't determined by any of the other variables in $\mathcal{U} \cup \mathcal{V}$. We can accomplish this within the nomic sufficiency account in the following way. Say that one model $\mathcal{M}' = (\mathcal{U}', \mathcal{V}', \mathcal{E}')$ *eclipses* another model $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$, $\mathcal{M} \sqsubset \mathcal{M}'$, iff $\mathcal{M}'$ and $\mathcal{M}$ share all the same variables and $\mathcal{M}'$ contains strictly more structural determination relations between those variables. That is:

$\mathcal{M} \sqsubset \mathcal{M}'$ iff:

1. $\mathcal{U} \cup \mathcal{V} \subseteq \mathcal{U}' \cup \mathcal{V}'$
2. $\forall V \in \mathcal{U} \cup \mathcal{V}, \quad \mathbf{PA}(V) \subseteq \mathbf{PA}'(V)$
3. $\exists V \in \mathcal{U} \cup V, \quad \mathbf{PA}(V) \subsetneq \mathbf{PA}'(V)$

(Where '$\mathbf{PA}'(V)$' is a vector of V's causal parents in the model $\mathcal{M}'$.) Now, we can enrich our account of the correctness of causal models by requiring that a model not be eclipsed by any other model which satisfies ($\mathcal{V}4$).

($\mathcal{M}2$) $\mathcal{M}$ is correct at ω iff:

1. $\mathcal{M}$ satisfies ($\mathcal{V}4$) at ω
2. $\neg\exists \mathcal{M}'$ such that $\mathcal{M}'$ satisfies ($\mathcal{V}4$) at ω and $\mathcal{M} \sqsubset \mathcal{M}'$

$$\mathcal{E}_9 = \begin{pmatrix} W := S \vee B \\ B := \overline{S} \end{pmatrix}$$

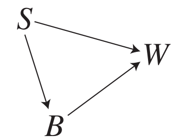


Figure 9

$$\mathcal{E}_{10} = \begin{pmatrix} W := S \vee B \\ S := \overline{B} \end{pmatrix}$$

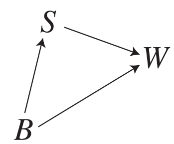


Figure 10

Returning to the example of Billy, Suzy, and the window (shown in figure 7): if Billy is eager to see the window shatter, and will throw his rock if (but only if) Suzy doesn't throw hers, then, given some assumptions about the selection function f, the system of structural equations shown in figure 9 will satisfy ($\mathcal{V}4$). (Here, '$\overline{*}$' is the truth function $1 - *$.) And this causal model eclipses the model consisting of the sole structural equation $W := S \vee B$, shown in figure 7. So, according to ($\mathcal{M}2$), the causal model in figure 7 will not be correct, if this one is. So, if Billy's decision about whether or not to throw is determined by whether Suzy throws, then the model consisting of just the equation $W := S \vee B$ is not correct. That model tells us that whether Billy throws isn't determined by whether Suzy throws, which is false. Similarly, if whether Suzy throws is determined by whether Billy does (she will throw iff he *doesn't*), then the system of structural equations shown in figure 10 will satisfy ($\mathcal{V}4$). And this causal model eclipses the one shown in figure 7. So that model will not be correct if this one is, according to ($\mathcal{M}2$).

3.4.2 Interventions

The nomic sufficiency account of causal models affords a specific and novel way of thinking about hypothetical interventions. On this understanding, a hypothetical intervention on an endogenous variable V is just a counterfactual supposition which takes one outside of the causal field $\mathfrak{F}_V$, while remaining inside the causal fields of all the other endogenous variables in the model. Consider, for instance, the causal model of the captain, the riflemen, and the deserter, reproduced below.
Suppose that this causal model is correct at world ω_0, and that $C_{\omega_0} = 0$—the cap-

$$\mathcal{E}_4 = \begin{pmatrix} L := C \\ R := C \\ D := L \vee R \end{pmatrix}$$

$$\begin{tikzcd} L \arrow[r] & D \\ C \arrow[u] \arrow[r] & R \arrow[u] \end{tikzcd}$$

Figure 4

tain doesn't give the order at ω_0. Suppose that we wish to entertain a hypothetical intervention to set the value of L to 1. We know that this is to be modeled in the following way: we take the original system of structural equations $\mathcal{E}_4$ and replace it with $\mathcal{E}_{4,\, L=1}$.

$$\mathcal{E}_{4,\, L=1} = \begin{pmatrix} L = 1 \\ R := C \\ D := L \vee R \end{pmatrix}$$

$$\begin{tikzcd} L \arrow[r] & D \\ C \arrow[r] & R \arrow[u] \end{tikzcd}$$

Figure 5

In the mutilated system of equations $\mathcal{E}_{4,\, L=1}$, L does not merely take on the value of 1. Additionally, the value of L is not structurally determined by the value of C. Whether the left rifleman fires is not determined by whether the captain gives the order. This follows immediately from the correctness of the system of equations $\mathcal{E}_{4,\, L=1}$, given ($\mathcal{M}2$), since if L were still structurally determined by C, then $\mathcal{M}_{4,\, L=1} =< (C), (L, R, D), \mathcal{E}_{4,\, L=1} >$ would be eclipsed by $\mathcal{M}_4 =< (C), (L, R, D), \mathcal{E}_4 >$. Since $\mathcal{M}_{4,\, L=1}$ is correct, it must not be so eclipsed, so L must not be structurally determined by C.

Since we've said that a structural equation ϕ_V is in force at a world ω iff ω lies inside of a causal field $\mathfrak{F}_V$ satisfying ($\mathfrak{F}2$), this means that a hypothetical intervention to set the value of L must take us to a world ω_1 which lies *outside* of $\mathfrak{F}_L$ (since L is not structurally determined by C), but still *inside* of $\mathfrak{F}_R$ and $\mathfrak{F}_D$ (since D is still structurally determined by L and R, and R is still structurally determined by C), as shown in figure 11(a). This provides a semantic interpretation of what's going on when we model an intervention on L by removing Ls's structural equation and leaving the other structural equations in place.

It also provides an explanation of why only certain methods of setting the value of L to 1 count as *interventions*, and it provides a criterion for distinguishing those ways of setting the values of the variables which do from those which do not constitute

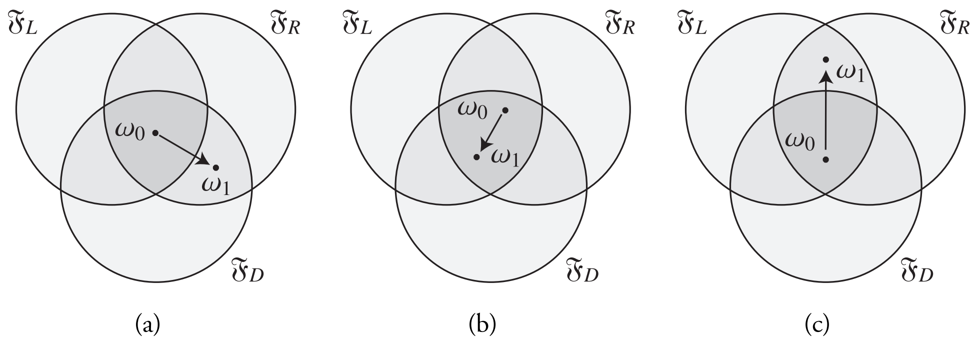


Figure 11

interventions. For instance, if we were to get the left rifleman to fire by bribing the captain to give him the order, then this would not constitute an intervention on L, since it would leave us *inside* of the causal field $\mathfrak{F}_L$, as shown in figure 11(b). Similarly, suppose that the captain does not want to kill the deserter, but would welcome an opportunity to let the riflemen blow off some steam. Then, we might be able to get the left rifleman to fire by putting up a bullet-proof partition between the riflemen and the deserter. Then, the captain would give the order, and the left rifleman would fire. Even though this is an intervention which makes the left rifleman fire, it is not an intervention *on* on the value of L. Rather, since it leaves L and R's determination by C intact, but severs the determination of D by L and R, as shown in figure 11(c), it constitutes an intervention on D.

This generalizes. A causal model $\mathcal{M}$ will be correct throughout $\mathfrak{F}_\mathcal{M} \stackrel{\text{def}}{=} \bigcap_{V \in \mathcal{V}} \mathfrak{F}_V$. This is the area of modal space in which every endogenous variable's structural equation $V \in \mathcal{V}$ is in force—it is the area in which all of the causal fields of the endogenous variables overlap. **Theorem 3**, proved in §3.6, establishes that, given ($\mathcal{M}2$), this area of modal space will contain every assignment of values to $\mathcal{U} \cup \mathcal{V}$ which is consistent with the structural equations in $\mathcal{E}$, and no assignments of values to $\mathcal{U} \cup \mathcal{V}$ which is not consistent with the structural equations in $\mathcal{E}$.

Theorem 3. *If a causal model $\mathcal{M} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ is correct according to ($\mathcal{M}2$), then $\mathfrak{F}_\mathcal{M} \stackrel{\text{def}}{=} \bigcap_{V \in \mathcal{V}} \mathfrak{F}_V$ contains all and only allowed assignment of values to the variables $V \in \mathcal{V}$, where an assignment is* allowed *just in case it is a solution to the equations in $\mathcal{E}$.*

This means that, for any assignment of values to the exogenous variables, there will be some area of modal space inside $\mathfrak{F}_\mathcal{M}$ where that assignment of values is realized.

So there are in principle hypothetical interventions to set the values of any of the exogenous variables without disrupting any of the structural determination relations in $\mathcal{E}$.

Moreover, it follows from ($\mathcal{M}2$) that for any correct causal model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$, for any $\mathbf{V} \subsetneq \mathcal{V}$, and any assignment of values to $\mathbf{V}$, there will always be an area of modal space which is *outside* the causal fields of all the members of $\mathbf{V}$ but still *inside* the causal fields of all of the members of $\mathcal{V} - \mathbf{V}$, and which contains every possible assignment of values to $\mathbf{V} \cup \mathcal{U}$.

Theorem 4. *Given* ($\mathcal{M}2$)*, for any* $\mathbf{V} \subsetneq \mathcal{V}$*,*

$$\bigcap_{W \notin \mathbf{V}} \mathfrak{F}_W - \bigcup_{V \in \mathbf{V}} \mathfrak{F}_V$$

is non-empty and contains every assignment of values to the variables in $\mathbf{V} \cup \mathcal{U}$*.*

This means that, if a structural equations model is correct, according to ($\mathcal{M}2$), then there is an in-principle intervention to set any subset of the variables in $\mathcal{V}$ to any assignment of values which will leave the structural equations of the non-intervened upon variables intact. This, together with **theorem 2** and **theorem 3**, guarantees that the structural equations in $\mathcal{E}$ are strongly modular.

3.4.3 Causal Fields and Selection Functions

In the preceding, I took the selection function f as primitive, and defined a condition on the causal field $\mathfrak{F}_V$ in terms of it. However, once we have the notion of a causal field, we needn't continue to take the selection function as primitive. If we think that we can provide an independent characterization of the notion of a causal field which will guarantee the truth of **theorems 2–4**, then we can use these causal fields to define the selection function f. We can just say that, relative to a correct causal model $\mathcal{M}$, at any world $\omega_0 \in \mathfrak{F}_{\mathcal{M}}$, $f(\mathbf{V} = \mathbf{v}, \omega_0)$ is the set of worlds outside of the causal fields of all the endogenous $V \in \mathbf{V}$ but still inside of the causal fields of all of the other variables in $\mathcal{V}$. That is, relative to a correct causal model $\mathcal{M}$ we may define $f(\mathbf{V} = \mathbf{v}, \omega_0)$ as follows.

$$f(\mathbf{V} = \mathbf{v}, \omega_0) \stackrel{\text{def}}{=} \left\{ \omega : \omega \notin \bigcup_{V \in \mathbf{V}} \mathfrak{F}_V \wedge \omega \in \bigcap_{W \notin \mathbf{V}} \mathfrak{F}_W \wedge \mathbf{V}_\omega = \mathbf{v} \right\}$$

Such a theory would be in line with the theory of Kment (2006), according to which the world's causal structure determines the comparative similarity relations between worlds used to evaluate causal counterfactuals.[34]

It is noteworthy that a theory like that has an easy time explaining the truth of ($f3$). Recall, the causal counterfactual account did not appear to have a story to tell about why the set of worlds that we consider when evaluating the hypothetical supposition that the left rifleman fired were worlds at which the left rifleman's decision to fire didn't depend upon the captain's order, but at which the deserter's death still depended upon whether the left rifleman fired. If we just define the selection function f so that it picks out the worlds that constitute *interventions*, in our technical sense, then ($f3$) is no longer mysterious; rather, it is a straightforward consequence of the definition of f together with our definition of *intervention*.

A theory along these lines owes an independent account of which worlds get included in the causal field $\mathfrak{F}_V$. I'm optimistic that such an account can be provided, but I won't attempt to provide it here. If the reader is less optimistic than I, then they are free to continue to take the selection function f as more primitive than the causal field $\mathfrak{F}_V$ in what follows.

3.4.4 Causal Counterfactual Dependence

($\mathcal{M}2$) allows us to provide an account of causal counterfactual conditionals in terms of relations of structural determination. On this account, a causal counterfactual $\mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C}$ is true at a world ω iff there is a correct structural equations model $\mathcal{M}$ such that, given the exogenous variable assignment $\mathcal{U}_\omega$, if $\mathcal{M}$ is minimally mutilated so as to make $\mathscr{A}$ true, then $\mathscr{C}$ is true in the resulting model.[35]

$$(\mathrel{\Box\!\!\rightarrow}_{\mathcal{M}}) \qquad \mathscr{A} \mathrel{\Box\!\!\rightarrow} \mathscr{C} \iff \mathcal{M}_{\mathscr{A}}, \mathcal{U}_\omega \models \mathscr{C}$$

With this kind of account, we could take structural determination to be more primitive than causal counterfactual dependence, and use the former to provide a reductive

[34] Kment (2006)'s account is slightly more general than this. He thinks that relations of *explanatory* relevance underlie the similarity relation. In the case of causal counterfactuals, however, the causal theory of explanation entails that it is at least in part *causal* relevance which determines similarity.

[35] I mean to leave it open whether the counterfactual is evaluated only relative to a correct model $\mathcal{M}$, or whether the counterfactual is evaluated by quantifying over all correct models, or whether context in some way selects a class of structural equations models relevant to evaluating the counterfactual. It is my view that context plays a large role in selecting which structural equations models are relevant to the evaluation of a causal claim (whose truth conditions are given by causal counterfactuals), but a defense of that claim will have to wait for another occasion.

$$\mathcal{E}_{12} = (W := B + B \cdot H)$$

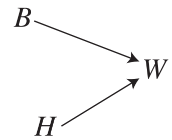


Figure 12

account of the latter. That is to say: with this account, we need not define causal counterfactual dependence directly in terms of the selection function; rather, f can be used to provide the truth conditions for $\mathcal{M}$, which can be used to provide the truth conditions for $\Box\!\!\rightarrow$. (Or, if we choose to define the selection function in terms of antecedently understood causal fields, as discussed in §3.4.3 above, then we could take facts about causal fields to ground both facts about f and facts about $\mathcal{M}$.)

The causal counterfactual understanding, ($\mathcal{M}1$), in contrast, retained an account of causal counterfactual conditionals according to which $\mathscr{A} \Box\!\!\rightarrow \mathscr{C}$ is true at the world of evaluation, ω, iff all the worlds in $f(\mathscr{A}, \omega)$ are worlds at which $\mathscr{C}$ is true.

$$(\Box\!\!\rightarrow_f) \qquad\qquad \mathscr{A} \Box\!\!\rightarrow \mathscr{C} \iff f(\mathscr{A}, \omega) \models \mathscr{C}$$

Depending upon our semantics for f, there may be cases in which ($\Box\!\!\rightarrow_{\mathcal{M}}$) and ($\Box\!\!\rightarrow_f$) diverge. Just to fix ideas: consider an account roughly like that of Lewis (1979b) or Maudlin (2007).[36] On Maudlin (2007)'s account, $f(\mathscr{A}, \omega)$ is the set of worlds that you get by performing a surgical alteration to ω so as to make $\mathscr{A}$ true at the relevant time, and then time-evolving the resulting state of the world forward in time according to the fundamental laws of nature. While Lewis (1979b)'s account is slightly more complicated, it will achieve the same results as Maudlin's in the case I'll be considering.

Imagine that I've got a tychistically chancy coin—whether it lands heads is not determined by the previous microphysical state of the universe and the laws of nature; rather, the previous state of the universe and the laws of nature assign a precise probability of one half to the coin landing heads and a probability of one half to the coin landing tails. I'm going to flip the coin, and I offer you a bet on whether or not the coin lands heads. I'm an honest player, so if you take the bet and the coin lands

[36] Again, Lewis's theory will not take the selection function as primitive, but for any true counterfactual $\mathscr{A} \Box\!\!\rightarrow \mathscr{C}$, we can define a set $f(\mathscr{A}, \omega)$ of closest $\mathscr{A}$-worlds which will entail that $\mathscr{C}$—see fn 22. Given the semantics of Lewis (1979b), this set will be a set of worlds identical to the actual world up until a time shortly before the antecedent, at which point, there will be a tiny miracle so as to bring about the truth of the antecedent.

heads, then you'll win some money. If you take the bet and the coin lands tails, then you'll lose some money. If you don't take the bet, then you'll neither win nor lose any money, independent of whether or not the coin lands heads. Let's stipulate that the chance that the coin lands heads is unaffected by whether you take the bet. In this scenario, it appears that the structural equations model shown in figure 12 is correct, where B is a binary variable that takes the value 1 if you accept the bet and 0 if you don't accept the bet, H is a binary variable which takes the value 1 if the coin lands heads and 0 if the coin lands tails, and W is a ternary variable which takes the value 0 if you neither win nor lose money, 1 if you lose money, and 2 if you win money. The structural equation says that whether you win is determined by whether you play and whether the coin lands heads. If you don't accept the bet, then you'll neither win nor lose; if you take the bet and the coin lands tails, then you'll lose; and if you take the bet and the coin lands heads, then you'll win. Let's say that, at the actual world, you refuse the bet, and the coin lands heads.

Suppose that we adopt the Maudlin account of the selection function. Then, ($\mathcal{M}1$) and ($\square\!\!\rightarrow_f$) will tell us that this structural equations model is not correct, since condition ($\mathcal{U}1$) will not be satisfied. ($\mathcal{U}1$), recall, required that, were some of the exogenous variables to have taken on different values, the other exogenous variables would have retained their actual values. However, $f(B = 1, @) \not\models H = 1$, since when we surgically alter the state of the world so as to make $B = 1$ true and time-evolve the resulting state of the world into the future according to fundamental the laws of nature, there are two possibilities: one in which the coin lands heads and one in which the coin lands tails. In contrast, the structural equations model $\mathcal{M}_{12} =< (B, H), (W), \mathcal{E}_{12} >$ will satisfy the second clause of ($\mathcal{M}2$)—($\mathcal{M}2$, 2)—so long as there is no other structural equations model which satisfies ($\mathcal{M}2$, 1) according to which W is determined by B and H, and either B is determined by H or H is determined by B. Since $f(B = 1, @)$ contains both $H = 0$ worlds and $H = 1$ worlds, there will be no set of worlds containing $f(B = 1, @)$ which entails that H is any function of B. So H is not determined by B, according to ($\mathcal{M}2$, 1). Assuming that, under the counterfactual supposition that the coin lands tails, you still refused the bet—*i.e.*, that $f(H = 0, @) \models B = 0$—$\mathcal{M}$ will satisfy ($\mathcal{M}2$, 2).

Of course, there's no reason that the counterfactual account can't replace ($\mathcal{U}1$) with the requirement that a causal model be uneclipsed. On an account like this, a model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ will be correct iff $\mathcal{V}$ satisfies ($\mathcal{V}1$) and there is no other structural equations model $\mathcal{M}'$ which both satisfies ($\mathcal{V}1$) and eclipses $\mathcal{M}$. Even this

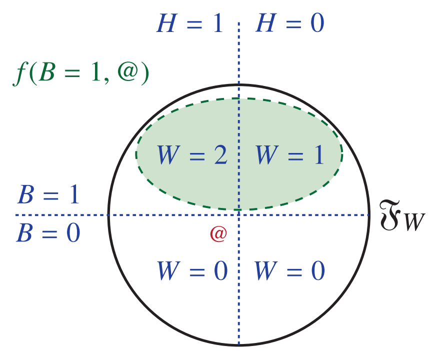


Figure 13: The relationship between the causal field $\mathfrak{F}_W$, the selection function f, and the causal counterfactual $B = 1 \;\Box\!\!\rightarrow\; W = 2$. According to ($\mathcal{M}2$) and ($\Box\!\!\rightarrow_{\mathcal{M}}$), the modal profile shown above is, in the presence of other plausible assumptions, sufficient for the truth of $B = 1 \;\Box\!\!\rightarrow\; W = 2$. However, according to ($\Box\!\!\rightarrow_f$), this modal profile is incompatible with the truth of $B = 1 \;\Box\!\!\rightarrow\; W = 2$, since $f(B = 1, @) \not\models W = 2$.

emended counterfactual account will fail to say that the system of equations in figure 12 is correct. That's because, even though the following counterfactuals will all be deemed true,

$$B = 0 \wedge H = 0 \;\Box\!\!\rightarrow\; W = 0$$

$$B = 1 \wedge H = 0 \;\Box\!\!\rightarrow\; W = 1$$

$$B = 0 \wedge H = 1 \;\Box\!\!\rightarrow\; W = 0$$

$$B = 1 \wedge H = 1 \;\Box\!\!\rightarrow\; W = 2$$

at the actual world (where, recall, you turned down the bet and the coin landed heads) the counterfactual $B = 1 \;\Box\!\!\rightarrow\; W = 2$ will be false. For $f(B = 1, @) \not\models W = 2$, since there are some worlds in $f(B = 1, @)$ where the coin lands tails and you therefore lose the bet ($W = 1$). So, ($\mathcal{V}1$), wedded with a Maudlin-esque account of the selection function, entails that the system of structural equations in figure 12 is incorrect.[37]

Independent of its ability to vindicate the system of structural equations $\mathcal{E}_{12}$, the fact that this account of the selection function, together with ($\Box\!\!\rightarrow_f$), entails the falsity of $B = 1 \;\Box\!\!\rightarrow\; W = 2$ strikes me, as it has struck many,[38] as the wrong result. Whether

[37] Some readers will feel tempted at this point to give up on the subvector clause of ($\mathcal{V}1$), and merely require that for every assignment of values $\overline{\mathbf{v}}$ to $\overline{\mathbf{V}}$, were those values realized, V would be $\phi_V(\mathbf{pa}_{\overline{\mathbf{v}}}(V))$. Those readers should reconsider the problems with transitivity from page 74. Note that, without the subvector clause, we will not be able to infer $C = 1 \;\Box\!\!\rightarrow\; D = 1$ from the correctness of the structural equations model shown in figure 6.

[38] See Bennett (2003, ch. 15) and Kment (2006)

the coin lands heads is entirely unaffected by whether you took the bet. Since the coin actually landed heads, if you had taken the bet, you *would* have won. $B = 1 \boxright W = 2$ is true. Now, there are moves to be pulled here—we can alter our account of the selection function so that, if the coin actually lands heads, then only the worlds where the coin lands heads are included in $f(B = 1, @)$. Note, however, that the account consisting of ($\mathcal{M}2$) and ($\boxright_{\mathcal{M}}$) need not avail itself of those maneuvers. Even with the bare MAUDLIN account of the selection function, that account entails that, were you to take the bet, you would have won. That's because ($\mathcal{M}2$) only requires that $f(B = 1, @)$ be included in $\mathfrak{F}_W$, and that $\mathfrak{F}_W \models W = B + B \cdot H$. The correctness of the structural equations model in figure 12 does not depend upon whether the worlds in $f(B = 1, @)$ are worlds in which the coin lands heads or tails, or whether they are worlds in which you win or lose. ($\mathcal{M}2$) only requires that, at all the worlds in $f(B = 1, @)$ at which the coin lands tails, you lose; and that, at all the worlds in $f(B = 1, @)$ at which the coin lands heads, you win. Assuming that similar remarks apply to all the other worlds in $\mathfrak{F}_W$, the structural determination relations shown in figure 12 will be in force. Then, those structural determination relations will entail, *via* ($\boxright_{\mathcal{M}}$), that $B = 1 \boxright W = 2$. So, according to ($\mathcal{M}2$) and ($\boxright_{\mathcal{M}}$), $B = 1 \boxright W = 2$ can be true even though $f(B = 1, @) \not\models W = 2$. Thus, according to ($\mathcal{M}2$) and ($\boxright_{\mathcal{M}}$), it is possible for a counterfactual $\mathscr{A} \boxright \mathscr{C}$ to be true at a world ω, even though $f(\mathscr{A}, \omega) \not\models \mathscr{C}$. (See figure 13.)

This is not easily mimicked by the causal counterfactual account, for that account is committed to both ($\boxright_{\mathcal{M}}$) and ($\boxright_f$). For instance, the counterfactual theorist might want to attempt to adopt the nomic sufficiency account's treatment of the coin toss case by emending ($\mathcal{V}1$) to read:

$$(\mathcal{V}5) \qquad \forall V \in \mathcal{V} \quad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \quad (\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \boxright V = \phi_V(\mathbf{PA}(V)))$$

($\mathcal{V}5$), unlike ($\mathcal{V}1$), does not require that, were $\overline{\mathbf{V}}^*$ to take on the values $\overline{\mathbf{v}}^*$, V would take on the value it is given in the mutilated model $\mathcal{M}_{\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*}$, with the actual assignment of values to the exogenous variables. It simply requires that, whatever values V's parent variables end up taking on when the values of $\overline{\mathbf{V}}^*$ change, the value of V remains a function ϕ_V of those values. This would allow the counterfactual theorist to agree with the nomic sufficiency theorist that were you to have taken the bet, you would have won, $B = 1 \boxright W = 2$. However, since the counterfactual theorist is still committed to ($\boxright_f$), so long as they retain the simple MAUDLIN account of the selections function, they must also *deny* that were you to have taken the bet, you would have won, $\neg(B = 1 \mathrel{\Box\!\!\rightarrow} W = 2)$, since $f(B = 1, @) \not\models W = 2$. This appears to be a straightforward contradiction. Of course, English counterfactuals are highly context-sensitive, and the English 'If $\mathscr{A}$ were the case, $\mathscr{C}$ would be the case' can be true in one context and false in another. However, this kind of context-sensitivity is usually handled by letting context have a role to play in determining f. I also think that context has a role to play in bringing to salience certain systems of structural determination. However, neither of these kinds of context-sensitivity are going to help keep the envisaged counterfactual account out of trouble. *Holding fixed* the selection function, and *holding fixed* the relations of structural determination, this account will still end up entailing $(B = 1 \mathrel{\Box\!\!\rightarrow} W = 2) \wedge \neg(B = 1 \mathrel{\Box\!\!\rightarrow} W = 2)$.

The causal counterfactual theorist might want to respond to these kinds of considerations by denying ($\mathrel{\Box\!\!\rightarrow}_f$), and reformulating their account of the correctness conditions of a structural equations model directly in terms of the selection function f, saying nothing of causal counterfactuals. That is, they could replace ($\mathcal{V}5$) with ($\mathcal{V}6$)

$$(\mathcal{V}6) \qquad \forall V \in \mathcal{V} \quad \forall \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \quad \forall \overline{\mathbf{v}}^* \quad f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, @) \models V = \phi_V(\mathbf{PA}(V))$$

It follows immediately from **theorem 1** that, in the presence of ($\mathcal{U}1$), ($\mathcal{V}6$) is equivalent to ($\mathcal{V}1$). Of course, we just encountered reason for such a theorist to abandon ($\mathcal{U}1$)—namely that, together with the Maudlin-esque account of f, it is inconsistent with the correctness of $\mathcal{E}_{12}$. And once ($\mathcal{U}1$) has been replaced with a condition along the lines of ($\mathcal{M}2$, 2), ($\mathcal{V}6$) will no longer be equivalent to ($\mathcal{V}1$).[39] Given that it denies any direct connection between f and counterfactual conditionals, we might well wonder whether the resulting account deserves the name 'counterfactual' any longer, but put that question to the side. Whatever we call the resulting account, it is only able to avoid complicating its account of f by inching ever closer to the nomic sufficiency account. The only thing separating the two accounts at this point is the nomic sufficiency account's closure condition, ($\mathfrak{F}2$). This is the aspect of the account which solves the problems with modularity raised in §3.3.1 above. Moreover, it appears that any counterfactual account built around ($\mathcal{V}6$) which was able to solve those problems with modularity would end up being equivalent to (or would entail) the nomic sufficiency account. For it appears that the only way to solve those problems is to impose a constraint on which worlds are reachable by repeated counterfactual supposition; for worlds that are so reachable, put them in the set $\mathfrak{F}_V$, and strong modularity will then

[39] The case currently under discussion provides a counterexample to the equivalence.

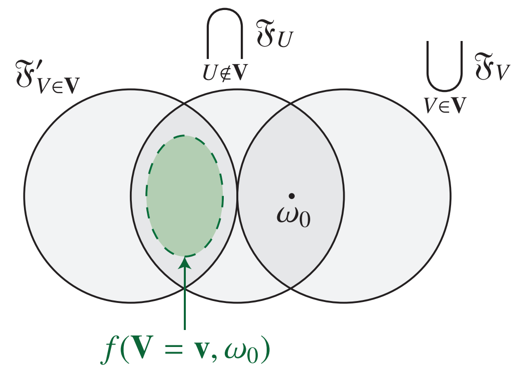


Figure 14: An intervened-upon variable may still depend upon its structural parents. Suppose that $\omega_0 \in \mathfrak{F}_{\mathcal{M}}$ and that, for each $V \in \mathbf{V}$, $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$. Then, the nomic sufficiency account fails to guarantee that $f(\mathbf{V} = \mathbf{v}, \omega_0)$ will not lie within *another* causal field $\mathfrak{F}'_V \models V = \phi'_V(\mathbf{PA}(V), \ldots)$, for some $V \in \mathbf{V}$.

guarantee that $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$. If that's right, then such an account would impose all the same constraints as the nomic sufficiency account; and counterfactual theorists would have mimicked nomic sufficiency theorists only by becoming nomic sufficiency theorists themselves.

3.4.5 A Remaining Worry

Above, I defined modularity as the thesis that any number of interventions on a set of variables $\mathbf{V}$ leaves the structural equations associated with every variable $V \notin \mathbf{V}$ unaffected. **Theorems 2–4** guarantee that a correct system of structural equations will be modular in this sense. Note, however, that modularity does not guarantee that there will always be an intervention on a set of variables $\mathbf{V}$ such that, post-intervention, the variables in $\mathbf{V}$ are no longer determined by any of the variables in $\mathbf{PA}(\mathbf{V})$. For a schematic example, consider the causal fields shown in figure 14. In that diagram, the causal model $\mathcal{M}$ will be correct at the world $\omega_0 \in \mathfrak{F}_{\mathcal{M}} \stackrel{\text{def}}{=} \bigcap_{V\in\mathcal{V}} \mathfrak{F}_V$. Now, **theorem 3** guarantees us that there is a set of worlds $f(\mathbf{V} = \mathbf{v}, \omega_0)$ which lie outside of $\bigcup_{V\in\mathbf{V}} \mathfrak{F}_V$, yet inside of $\bigcap_{U\notin\mathbf{V}} \mathfrak{F}_U$, and at which $\mathbf{V} = \mathbf{v}$, for any assignment $\mathbf{v}$. However, we have no guarantee that this set of worlds *doesn't* lie within some *other* causal field $\mathfrak{F}'_V$ such that $\mathfrak{F}'_V \models V = \phi'_V(\mathbf{PA}(V), \ldots)$, for some $V \in \mathbf{V}$. And that means that, even though we have a guarantee that an *intervention* on a set of variables will sever the actual structural determination relations between $\mathbf{V}$ and $\mathbf{PA}(\mathbf{V})$, we *don't* have any guarantee that the intervention won't make it the case that some *other* structural determination relations link $\mathbf{PA}(\mathbf{V})$ to $\mathbf{V}$.

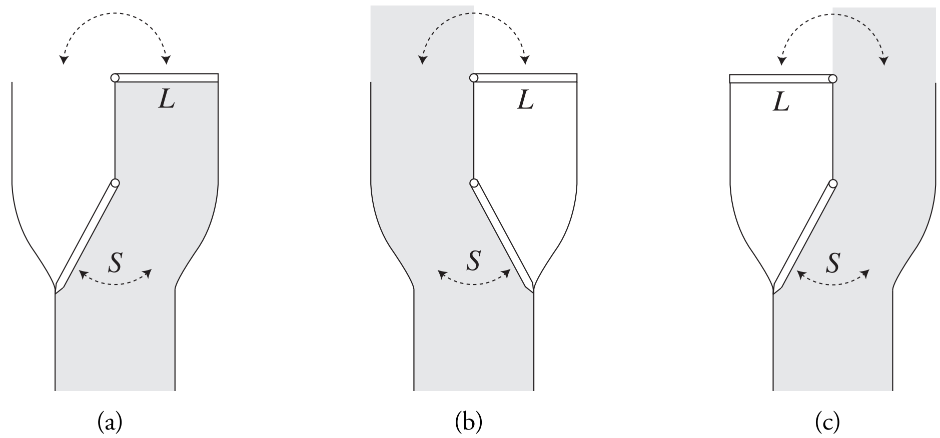


Figure 15

For a concrete example which might give rise to a case like this, consider the steam vent illustrated in figure 15. There, a switch, which may be placed to the left or to the right, will either divert the steam to the left or the right. (If the switch is left, the steam will go right, as shown in figure 15(a); if the switch is right, the steam will go left, as shown in figure 15(b).) There is a lid on the right steam vent. If the steam is directed up to the right vent, then the lid will heat up. Consider the variables S and L. S is 1 if the switch is to the left, and is 0 if the switch is to the right. L is 1 if the lid is hot and is 0 if the lid is not hot. When the system is as depicted in figure 15(a), the structural equation $L := S$ will be in force. Whether the lid is hot is determined by whether the switch is to the left or right. In figure 15(a), both S and L will be 1. Now, suppose that the lid is attached to a hinge, so that it can be pivoted to sit atop either the left or the right steam vent. There is then an intervention we may perform to set L to 0. That is, there is a method for making the lid not hot which will take us outside of the causal field $\mathfrak{F}_L$. We may simply pivot the lid on its hinge to put it atop the left steam vent, as in figure 15(c). Then, it will no longer be the case that $L = 1$, nor will it be the case that S determines L according to the equation $L := S$. However, even after this intervention has taken place, the value of S will determine the value of L. It will now do so according to the equation $L := \overline{S}$. If the switch is set to the left, then the lid will not be hot, and if the switch is set to the right, then the lid will be hot.

Given the account of interventions provided in §3.4.2 above, this will count as an intervention on the value of L. However, it would be inappropriate to model the result of this intervention by mutilating the model, removing L's structural equation, and replacing it with nothing. For, in order for the mutilated model to be correct, it must be uneclipsed by any correct structural equations model. And, in this case, the mutiliated model in which S does not structurally determine L would be eclipsed by the model containing the equation $L := \overline{S}$.

Cases such as these might also make trouble for ($\Box\!\!\rightarrow_{\mathcal{M}}$), as ($\Box\!\!\rightarrow_{\mathcal{M}}$) would predict that, if we are in the situation depicted in figure 15(a), then, were the lid to not be hot, then, if the switch were moved to the right, the lid would not be hot, *i.e.*, $L = 0 \Box\!\!\rightarrow (S = 0 \Box\!\!\rightarrow L = 0)$. However, if we think that there are worlds in $f(L = 0, @)$ at which the lid has been pivoted on its hinge, then we might think that this counterfactual should be false; rather, we may think, were the lid to not be hot, then, if the switch were moved to the right, the lid *would* be hot—*i.e.*, $L = 0 \Box\!\!\rightarrow (S = 0 \Box\!\!\rightarrow L = 1)$. Or perhaps, were the lid to not be hot, then, if the switch were moved to the right, the lid *might* be hot—*i.e.*, $L = 0 \Box\!\!\rightarrow (S = 0 \Diamond\!\!\rightarrow L = 1)$.[40]

It is unclear to me whether this ought to be regarded as a problem for the nomic sufficiency account. To the extent that one is inclined to think that $f(L = 0, @)$ includes worlds at which the lid has been pivoted on its hinge, it seems to me entirely correct to say that we ought not model an intervention on L which pivots the lid on its hinge by removing the structural determination relation between L and S, and it seems a mark in the nomic sufficiency account's favor that it says so. To the extent that one is inclined to think that $f(L = 0, @)$ contains worlds at which the lid is removed from its hinge, or perhaps worlds at which some kind of Lewisian miracle keeps the lid from getting hot even though the steam is being directed up towards it, it seems entirely correct to model this kind of intervention by removing the structural determination relation between S and L. Once there is a miracle to set L to 0, changes in the value of S will not affect the value of L, so long as God's hand is steady. If one is unhappy with the possibility of L being determined by S post-intervention, then one may simply require that $f(L = 0, @)$ contain only worlds

[40] I haven't said anything about how to understand so-called might counterfactuals, of course. One might simply define them as the dual of the standard counterfactual $\mathscr{A} \Diamond\!\!\rightarrow \mathscr{C} \stackrel{\text{def}}{=} \neg(\mathscr{A} \Box\!\!\rightarrow \neg\mathscr{C})$. However, I don't think that we should; in part because I think that this duality leads pretty quickly into counterfactual skepticism—the view that almost all counterfactuals are false. (See DeRose, 1999, Hawthorne, 2005, and Hájek, ms.) At the moment, I'd rather say that might counterfactuals are epistemic modals wide-scoping over would counterfactuals—*i.e.*, $\mathscr{A} \Diamond\!\!\rightarrow \mathscr{C} \stackrel{\text{def}}{=} \Diamond(\mathscr{A} \Box\!\!\rightarrow \mathscr{C})$. But I don't think that how we come down on these issues makes any difference to the issues being discussed in the body.

at which LEWISIAN miracles determine the value of L. Of course, nothing in the account guarantees that there will always be some possible LEWISIAN miracle which will constitute an intervention. However, if there are cases in which LEWISIAN miracle interventions are impossible, then I'm inclined to say just what I said above about the case in which $f(L = 0, @)$ contained worlds at which the lid was pivoted on its hinge: in such cases, it is incorrect to model the result of the intervention by mutilating the model, and it would be a mark against an account of structural determination if said otherwise. (Of course, we might want to use the term *intervention* in such a way that these ways of setting variable values don't constitute interventions. That's a matter of semantic legislation; I'd prefer to count them as interventions, but I don't see how anything of substance hangs on it.) So my settled judgment is that this is a feature, rather than a bug, of the nomic sufficiency account, though I wouldn't be too surprised to be persuaded otherwise.

3.5 In Summation

I developed and explored a common way of understanding structural equations models, according to which they represent patterns of causal counterfactual dependence between variable values. I argued that it faces difficulties in securing the independent manipulability—the *modularity*—of the structural determination relations codified in a correct system of structural equations. I put forward a different way of understanding structural equations models, according to which one variable value is structurally determined by others only if the values of the latter variables are nomically sufficient for the value of the former throughout a certain region of modal space. I showed how this way of understanding structural equations models allows us to secure the independent manipulability of structural determination relations. I showed that it provides a clear and straightforward way of thinking about hypothetical interventions, as well as a criterion for distinguishing hypothetical changes in the values of variables which constitute interventions from those that do not. Finally, I showed that, since the account does not require causal counterfactual conditionals to undergird the correctness of a structural equations model, it is free to use structural equations models to provide a reductive account of causal counterfactuals. By treating structural determination relations as more fundamental than causal counterfactuals, the resulting theory was able to yield a clean solution to a problem case that has troubled previous accounts of causal counterfactuals.

3.5.1 Looking Forward

In chapter 5, I will provide an account of singular causation in terms of the networks of structural determination represented in a correct structural equations model. That account will be very much in the spirit of Lewis's 1973 counterfactual account. I will say that the event of one variable C's having or changing its value at a time caused the event of another variable E's having or changing its value at a time if and only if the later event counterfactually depends—in at least one way—upon the former within a correct structural equations model, *or* there is a certain kind of chain of counterfactual dependence running from the latter event to the former. By carefully fleshing out the details, we will see that this account can avoid the plethora of counterexamples which have emerged in response to Lewis's original 1973 account.

However, before getting to that, there is a serious problem to be dealt with. The problem, which will be developed at length in chapter 4, is that a correct structural equations model—on either the counterfactual or the nomic sufficiency understanding—underdetermines singular causal relations. That is, there exist pairs of systems which are, according to both the counterfactual and the nomic sufficiency account, correctly modeled by the same structural equations models, but which nevertheless differ causally. In one system, two variable's values are causally related; whereas, in the second system, the corresponding variable's values are *not* causally related. We must therefore say more about the conditions under which a correct structural equations model may be used to infer something about which events singularly caused with other events. The conclusion of chapter 4 will be that we may do so iff all of the structural determination relations in the model are *autonomous*, in a sense to be made precise below.

3.6 Proofs

Define the *rank* of a variable $V \in \mathcal{U} \cup \mathcal{V}$ recursively as follows:

$$
\begin{aligned}
rank(V) = 0 &\quad\Longleftrightarrow\quad V \in \mathcal{U} \\
rank(V) = k+1 &\quad\Longleftrightarrow\quad \max\{rank(U) \;:\; U \in \mathbf{PA}(V)\} = k
\end{aligned}
$$

Graphically, a variable's rank is the largest number of edges lying between that variable and an exogenous variable along a directed path. Let '$\mathbf{Rank}(i)$' denote the set of all variables of rank i, and let '$\mathbf{Rank}(i, j, ..., k)$' denote the union $\mathbf{Rank}(i) \cup \mathbf{Rank}(j) \cup ... \cup \mathbf{Rank}(k)$.

Lemma 1. *Given* ($\mathcal{V}1$), ($\mathcal{U}1$), *and* ($f1$), *for all* $V \in \mathcal{V}$, *all* $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, *all* $\overline{\mathbf{v}}^*$, *and all* $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$, *where* $\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$ *assigns the values to* $\mathbf{PA}(V)$ *determined by the structural equations in* $\mathcal{E} - \bigcup_i \{\phi_{U_i}\}$, *for every endogenous* $U_i \in \overline{\mathbf{V}}^*$, *and* $(\mathcal{U} - \overline{\mathbf{V}}^*)_\omega \cup \overline{\mathbf{v}}^*$.

Proof. By induction on the rank of the variables in $\mathcal{V}$.

Base Case. *For all* $V \in \mathbf{Rank}(1)$, *all* $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, *all* $\overline{\mathbf{v}}^*$, *and all* $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$.

Proof. If $rank(V) = 1$, then every $X \in \mathbf{PA}(V)$ is exogenous. Without loss of generality, consider one $X \in \mathbf{PA}(V)$. If $X \in \overline{\mathbf{V}}^*$, then $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_{\overline{\mathbf{v}}^*}$ (the value assigned to X by $\overline{\mathbf{v}}^*$), by ($f1$). If $X \notin \overline{\mathbf{V}}^*$, then $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_\omega$, by ($\mathcal{U}1$). In either case, X takes on the value assigned to it by $\mathbf{pa}_{\overline{\mathbf{v}}^*}$. □

Inductive Step. *If for all* $V \in \mathbf{Rank}(1, 2, ..., k)$, *it is true that, for all* $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, *all* $\overline{\mathbf{v}}^*$, *and all* $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$, *then for all* $V \in \mathbf{Rank}(k+1)$, *it will be true that, for all* $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, *all* $\overline{\mathbf{v}}^*$, *and all* $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$.

Proof. Without loss of generality, consider one $V \in \mathbf{Rank}(k+1)$, one $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, one $\overline{\mathbf{v}}^*$, and one $X \in \mathbf{PA}(V)$. Either $X \in \overline{\mathbf{V}}^*$ or $X \notin \overline{\mathbf{V}}^*$. Suppose that $X \in \overline{\mathbf{V}}^*$. Then, $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_{\overline{\mathbf{v}}^*}$, by ($f1$). If $X \notin \overline{\mathbf{V}}^*$, then, since $rank(X) \leq k$, $\mathbf{PA}(X)_{\omega'} = \mathbf{pa}(X)_{\overline{\mathbf{v}}^*}$ (the values of X's parents determined by $\mathcal{E} - \bigcup_i \{\phi_{U_i}\}$, for all $U_i \in \overline{\mathbf{V}}^*$, and $(\mathcal{U} - \overline{\mathbf{V}}^*)_\omega \cup \overline{\mathbf{v}}^*$), for all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, by the inductive hypothesis (since $X \notin \overline{\mathbf{V}}^*, \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{X}}$). Then, ($\mathcal{V}1$) guarantees that $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = \phi_X(\mathbf{pa}(X)_{\overline{\mathbf{v}}^*})$. So, whether $X \in \overline{\mathbf{V}}^*$ or $X \notin \overline{\mathbf{V}}^*$, X takes on the value $X_{\overline{\mathbf{v}}^*}$ at every

$\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$. Since X, V, $\overline{\mathbf{V}}^*$, and $\overline{\mathbf{v}}^*$ were arbitrary, for all $V \in \mathbf{Rank}(k+1)$, all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, and all $\overline{\mathbf{v}}^*$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}$ for every $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$. □

□

Lemma 2. *Given ($\mathcal{V}3$), ($\mathcal{U}1$), and ($f1$) for all $V \in \mathcal{V}$, all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, all $\overline{\mathbf{v}}^*$, and all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$, where $\mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$ assigns the values to $\mathbf{PA}(V)$ determined by the assignment of values $(\mathcal{U} - \overline{\mathbf{V}}^*)_\omega \cup \overline{\mathbf{v}}^*$ and the structural equations in $\mathcal{E} - \bigcup_i \{\phi_{U_i}\}$, for every endogenous $U_i \in \overline{\mathbf{V}}^*$.*

Proof. By induction on the rank of the variables in $\mathcal{V}$.

Base Case. *For all $V \in \mathbf{Rank}(1)$, all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, all $\overline{\mathbf{v}}^*$, and all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$.*

Proof. Consider, without loss of generality, a variable $V \in \mathbf{Rank}(1)$. Since V's rank is 1, every $X \in \mathbf{PA}(V)$ is exogenous. If $X \in \overline{\mathbf{V}}^*$, then $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_{\overline{\mathbf{v}}^*}$ (the value assigned to X by $\overline{\mathbf{v}}^*$), by ($f1$). If $X \notin \overline{\mathbf{V}}^*$, then $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_\omega$, by ($\mathcal{U}1$). In either case, X takes on the value assigned to it by $\mathbf{pa}_{\overline{\mathbf{v}}^*}$. □

Inductive Step. *If for all $V \in \mathbf{Rank}(1, 2, ..., k)$, it is true that, for all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, all $\overline{\mathbf{v}}^*$, and all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$, then for all $V \in \mathbf{Rank}(k+1)$, it will be true that, for all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, all $\overline{\mathbf{v}}^*$, and all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}(V)$.*

Proof. Without loss of generality, consider one $V \in \mathbf{Rank}(k+1)$, one $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, one $\overline{\mathbf{v}}^*$, and one $X \in \mathbf{PA}(V)$. Either $X \in \overline{\mathbf{V}}^*$ or $X \notin \overline{\mathbf{V}}^*$. Suppose that $X \in \overline{\mathbf{V}}^*$. Then, $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = X_{\overline{\mathbf{v}}^*}$, by ($f1$). If $X \notin \overline{\mathbf{V}}^*$, then, since $rank(X) \leq k$, $\mathbf{PA}(X)_{\omega'} = \mathbf{pa}(X)_{\overline{\mathbf{v}}^*}$ (the values of X's parents determined by $\mathcal{E} - \bigcup_i \{\phi_{U_i}\}$, for all $U_i \in \overline{\mathbf{V}}^*$, and $(\mathcal{U} - \overline{\mathbf{V}}^*)_\omega \cup \overline{\mathbf{v}}^*$), for all $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$, by the inductive hypothesis (since $X \notin \overline{\mathbf{V}}^*, \overline{\mathbf{V}}^* \subseteq \overline{\mathbf{X}}$). Then, ($\mathcal{V}3$) and ($\mathfrak{F}1$) guarantee that $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \subset \mathfrak{F}_X$ and $\mathfrak{F}_X \models X = \phi_X(\mathbf{pa}(X)_{\overline{\mathbf{v}}^*})$. So $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models X = \phi_X(\mathbf{pa}(X)_{\overline{\mathbf{v}}^*})$ So, whether $X \in \overline{\mathbf{V}}^*$ or $X \notin \overline{\mathbf{V}}^*$, X takes on the value $X_{\overline{\mathbf{v}}^*}$ at every $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$. Since X, V, $\overline{\mathbf{V}}^*$, and $\overline{\mathbf{v}}^*$ were arbitrary, for all $V \in \mathbf{Rank}(k+1)$, all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, and all $\overline{\mathbf{v}}^*$, $\mathbf{PA}(V)_{\omega'} = \mathbf{pa}_{\overline{\mathbf{v}}^*}$ for every $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$. □

□

Theorem 1. *In a causal model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$, if $\mathcal{U}$ satisfies ($\mathcal{U}1$), then $\mathcal{V}$ satisfies ($\mathcal{V}3$) iff $\mathcal{V}$ satisfies ($\mathcal{V}1$).*

Proof. First assume that $\mathcal{V}$ satisfies ($\mathcal{V}1$). Then, we know that for all $V \in \mathcal{V}$, all $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, and all assignments $\overline{\mathbf{v}}^*$ to $\overline{\mathbf{V}}^*$,

$$f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models V = \phi_V(\mathbf{pa}_{\overline{\mathbf{v}}^*}(V))$$

By **lemma 1**, it then follows that

$$\forall \omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega), V_{\omega'} = \phi_V(\mathbf{PA}(V)_{\omega'})$$

So

$$f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models V = \phi_V(\mathbf{PA}(V))$$

So, if for every V, every $\overline{\mathbf{V}}^*$, and every $\overline{\mathbf{v}}^*$, we include every $\omega' \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega)$ in $\mathfrak{F}_V$, then we will have a set $\mathfrak{F}_V$ which satisfies ($\mathfrak{F}1$) and which entails that $V = \phi_V(\mathbf{PA}(V))$. So every $V \in \mathcal{V}$ will satisfy ($\mathcal{V}3$).

To establish the other direction, assume that $\mathcal{V}$ satisfies ($\mathcal{V}3$). Then, for every $V \in \mathcal{V}$, every $\overline{\mathbf{V}}^*$, and every $\overline{\mathbf{v}}^*$, $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \in \mathfrak{F}_V$ and $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$. By **lemma 2**, it then follows that, for all V, $\overline{\mathbf{V}}^*$, and $\overline{\mathbf{v}}^*$,

$$f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega) \models V = \phi_V(\mathbf{pa}_{\overline{\mathbf{v}}^*}(V))$$

So $\mathcal{V}$ must satisfy ($\mathcal{V}1$) as well.

□

Theorem 2. *If ϕ_V belongs to a causal model satisfying ($\mathcal{V}4$) at a world ω_0, then ϕ_V will continue to belong to a causal model satisfying ($\mathcal{V}4$) after any number of consecutive hypothetical interventions to set the values of any $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \stackrel{\text{def}}{=} (\mathcal{U} \cup \mathcal{V}) - V$.*

Proof. By induction on the number of interventions.

Inductive Step. *If ϕ_V belongs to a causal model satisfying ($\mathcal{V}4$) at world ω_k after k interventions to set the values of any $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$, then ϕ_V will belong to a causal model satisfying ($\mathcal{V}4$) at the world ω_{k+1} where there is a $k+1^{st}$ intervention to set the values of any $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$.*

Proof. By the inductive hypothesis, ϕ_V belongs to a causal model satisfying ($\mathcal{V}4$) at ω_k. This means that there must exist a set of worlds $\mathfrak{F}_V$ which satisfies ($\mathfrak{F}2$) and which contains ω_k. By ($f3$), an intervention setting the value of some $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}}$ to $\overline{\mathbf{v}}^*$ must take us to a world $\omega_{k+1} \in f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega_k)$. Since $\omega_k \in \mathfrak{F}_V$, ($\mathfrak{F}2$)

guarantees that $f(\overline{\mathbf{V}}^* = \overline{\mathbf{v}}^*, \omega_k) \subseteq \mathfrak{F}_V$, so $\omega_{k+1} \in \mathfrak{F}_V$ as well. And, by assumption, $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$. So there is a $\mathfrak{F}_V \ni \omega_{k+1}$ such that $\mathfrak{F}_V$ satisfies ($\mathfrak{F}2$) and $\mathfrak{F}_V \models V = \phi_V(\mathbf{PA}(V))$. As ϕ_V was arbitrary, the same holds for every $V' \notin \overline{\mathbf{V}}^*$. So, ($\mathcal{V}4$) will be correct at ω_{k+1}. So ϕ_V will belong to a causal model satisfying ($\mathcal{V}4$) at ω_{k+1}. $\square$

Setting $k = 0$ in the proof of the inductive step establishes the base case. $\square$

Theorem 3. *If a causal model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ is correct according to ($\mathcal{M}2$), then $\mathfrak{F}_\mathcal{M} \stackrel{\text{def}}{=} \bigcap_{V \in \mathcal{V}} \mathfrak{F}_V$ contains all and only allowed assignment of values to the variables $V \in \mathcal{V}$, where an assignment is* allowed *just in case it is a solution to the equations in $\mathcal{E}$.*

Proof. The proof proceeds by induction on the rank of the variables in $\mathcal{V}$.

Base Case. *$\mathfrak{F}_\mathcal{M}$ contains all and only allowed assignment of values to the variables in* $\mathbf{Rank}(0)$.

Proof. For every $V \in \mathcal{V}$, $\mathfrak{F}_V$ contains $f(\mathcal{U} = \mathbf{u}, \omega)$, for every assignment $\mathbf{u}$ to $\mathcal{U}$, and every $\omega \in \mathfrak{F}_V$. So $\mathfrak{F}_\mathcal{M}$ contains $f(\mathcal{U} = \mathbf{u}, \omega)$, for every assignment $\mathbf{u}$ to $\mathcal{U}$ and every $\omega \in \mathfrak{F}_\mathcal{M}$. If $U \in \mathbf{Rank}(0)$, then U is exogenous, $U \in \mathcal{U}$. Every assignment of values to the exogenous variables is allowed. So $\mathfrak{F}_\mathcal{M}$ contains all and only allowed assignments to the variables in $\mathbf{Rank}(0)$. $\square$

Inductive Step. *If $\mathfrak{F}_\mathcal{M}$ contains all and only allowed assignment of values to the variables in* $\mathbf{Rank}(0, 1, ..., k)$, *then it contains all and only allowed assignment of values to the variables in* $\mathbf{Rank}(0, 1, ..., k, k+1)$.

Proof. Take an arbitrary $V \in \mathbf{Rank}(k+1)$. Since $\mathbf{PA}(V) \subseteq \mathbf{Rank}(0, 1, ..., k)$, the inductive hypothesis gets us that every and only the allowed assignment of values to $\mathbf{PA}(V)$ are realized in $\mathfrak{F}_\mathcal{M}$. And because $\mathcal{M}$ is correct, $V_\omega = \phi(\mathbf{PA}(V)_\omega)$, for every $\omega \in \mathfrak{F}_V$. Since $\mathfrak{F}_\mathcal{M} \subseteq \mathfrak{F}_V$, this means that $V_\omega = \phi(\mathbf{PA}(V)_\omega)$ for every $\omega \in \mathfrak{F}_\mathcal{M}$ as well. So $\mathfrak{F}_\mathcal{M}$ contains all and only the allowed values of V. Since V was arbitrary, the above holds for every $V \in \mathbf{Rank}(k+1)$. $\square$

$\square$

Theorem 4. *Given ($\mathcal{M}2$), for any $\mathbf{V} \subsetneq \mathcal{V}$,*

$$\bigcap_{W \notin \mathbf{V}} \mathfrak{F}_W - \bigcup_{V \in \mathbf{V}} \mathfrak{F}_V$$

is non-empty and contains every assignment of values to the variables in $\mathbf{V} \cup \mathcal{U}$.

Proof. Take an arbitrary $\mathbf{V} \subsetneq \mathcal{V}$, an arbitrary assignment of values $\mathbf{v}$ to $\mathbf{V}$, an arbitrary $W \notin \mathbf{V}$, and an arbitrary assignment $\mathbf{u}$ to $\mathcal{U}$. Then, $\mathfrak{F}_W$ contains worlds at which $\mathbf{V} \cup \mathcal{U}$ is set to $\mathbf{v} \cup \mathbf{u}$ by an intervention, by ($\mathfrak{F}2$) and ($f3$). Therefore, for every $\mathbf{u}$ and every $V \in \mathbf{V}$, there are worlds $\omega \in \mathfrak{F}_W$ at which $\mathcal{U}_\omega = \mathbf{u}$ and $V_\omega \neq \phi_V(\mathbf{PA}(V)_\omega)$. However, for every $V \in \mathbf{V}$, $\mathfrak{F}_V$ does *not* contain a world ω at which $V_\omega \neq \phi_V(\mathbf{PA}(V)_\omega)$. Since W, $\mathbf{V}$, $\mathbf{v}$, $\mathbf{u}$, and V were arbitrary, for every $W \notin \mathbf{V}$, every $\mathbf{u}$, and every $V \in \mathbf{V}$, there are worlds in $\mathfrak{F}_W$ which are not in $\mathfrak{F}_V$ and at which the value of $V \cup \mathcal{U}$ is set to any value $\mathbf{v} \cup \mathbf{u}$. Thus, $\bigcap_{W \notin \mathbf{V}} \mathfrak{F}_W - \bigcup_{V \in \mathbf{V}} \mathfrak{F}_V$ is non-empty and contains every assignment of values to the variables in $\mathbf{V} \cup \mathcal{U}$. $\square$

CHAPTER 4

THE UNDERDETERMINATION OF SINGULAR CAUSATION

4.1 Introduction

In the previous chapter, I introduced and provided an account of structural equations models. As I mentioned there, these models have been used, by various authors in various ways, to construct accounts of singular causation.[1] When undertaking a project like this, there are two distinct but interrelated questions to answer. Firstly, there is the question to which I dedicated the previous chapter—*viz.*, what does it take for one of these models to be correct? Secondly, there is the question to which I will dedicate chapter 5—*viz.*, what is the relationship between a correct structural equations model and relations of singular causation?

A broad consensus has begun to emerge about how to answer these two questions. As I mentioned in the previous chapter, with respect to the first question, the consensus appears to be that what it takes for an isolated structural equation

[1] See in particular Hitchcock (2001, 2007), Woodward (2003), Halpern & Pearl (2001, 2005), Hiddleston (2005b), Menzies (2007), Halpern (2008), and Halpern & Hitchcock (2010).

$V := \phi_V(\mathbf{PA}(V))$ to be correct is, *inter alia*, the truth of the causal counterfactuals 'were $\mathbf{PA}(V)$ to take on the values $\mathbf{pa}(V)$, V would take on the value $\phi_V(\mathbf{pa}(V))$', for each assignment $\mathbf{pa}(V)$ to the variables in $\mathbf{PA}(V)$. While these causal counterfactuals form the backbone of the account of the correctness conditions of structural equations models, most believe that there are other conditions that a structural equations model must meet in order to be correct. For instance, everybody seems to accept that the variables in $\mathcal{V}$ must be *distinct*, in the sense of Lewis (1986b). That is, there cannot be any purely logical relationship between those variables. So, for instance, $\mathcal{V}$ cannot contain both a variable for *rainfall in Ann Arbor* and *precipitation in Ann Arbor*, since the value of the latter (non-causally) puts an upper bound on the value of the former. Hall (2007) has an extended discussion of what kinds of variables can be allowed to figure in a structural equations model. Glymour et al. (2010) suggest that more attention must be paid to the ways in which variable values *change* over time. I'm going to ignore most of these additional constraints here, just as I ignored them in the previous chapter, because, with the exception of a few from Hitchcock (2007) and Halpern & Hitchcock (2010), none of these additional constraints are going to make a difference to my central thesis here (just as none of them would have made a difference to my arguments in the previous chapter). I'll discuss the exceptions in §4.5.

With respect to the second question, the consensus appears to be that, while the correctness of a system of equations within which $E \in \mathbf{DE}(C)$ is not sufficient for the event of a variable C taking on the value c to cause the event of a variable E taking on the value e, the correctness of such a system of equations is *necessary* for $C = c$ to cause $E = e$. These authors have attempted to provide additional conditions which, together with the correctness of the structural equation model, will be sufficient for $C = c$ causing $E = e$.

Some counterexamples to this general approach to answering the second question have emerged in the work of Hiddleston (2005b) and Hall (2007). These counterexamples demonstrate that an apparently correct structural equations model *underdetermines* the causal structure of the variables appearing in the equations. That is: if we accept the orthodox answer to the first question, then there will be a single structural equations model which will correctly model two different systems; yet, in one of these systems, the event represented by $C = c$ caused $E = e$; whereas, in the other, $C = c$ did not cause $E = e$. To deal with these counterexamples, some[2] have amended their accounts to include information about variables' *default*, *normal*, or *typical* values. Unfortunately, in this chapter, I must be the bearer of bad news. If we provide the consensus answer to the first question—that what it is for a system of structural equations to be correct is for some set of associated causal counterfactuals to be true—then there are further underdetermination worries which cannot be dealt with by including a normality ranking over variable values. That is, in this chapter I will demonstrate that information about singular causation cannot be gleaned from a correct systems of equations together with a specification of the actual variable values and a normality ranking of the variable's possible values, so long as what it takes for a model to be correct is spelled out in terms of causal counterfactuals. In fact, things are a bit worse than this. There are three *prima facie* plausible conditions on the correctness of a structural equations model, much weaker than the counterfactual account, which suffice on their own to get us the conclusion that singular causal relations cannot be inferred from a correct structural equations model, the actual variable values, and a normality ranking of those variable's possible values. So this is not a problem for the causal counterfactual account alone; indeed, the nomic sufficiency account outlined and defended in chapter 3 is also committed to these three principles. So, if I am to provide an account of singular causation formulated in term of systems of structural equations—as I hope to do in chapter 5—then I must say more about when we can, and when we cannot, infer singular causal facts from a correct structural equations model.

In §4.2, I will introduce one of the more prominent attempts to define singular causation in terms of a correct structural equations model—the account of Halpern & Pearl (2005)—and outline some counterexamples to that account, as well as the revision of Halpern (2008) which appeals to the *normality* of a variable's potential values. After presenting some problems for Halpern (2008)'s normality-based approach in §4.3, I will go on in section §4.4 to demonstrate that, given three principles which follow from, but are strictly weaker than, the causal counterfactual understanding of structural equations models, the correctness of a structural equations model, together with a normality ranking on the variable's potential values, underdetermines the singular causal relations holding between the variables appearing in that model. In section §4.5, I'll consider some objections and outline some potential strategies for

[2] Hall (2007), Hitchcock (2007), and Halpern (2008). See also Hitchcock & Knobe (2009).

coping with this bad news. I'll ultimately decide that the relations of structural determination appearing in a correct structural equations model must be *autonomous*—in a sense to be made precise below—if we are to infer singular causal facts from the model.

4.2 The Halpern-Pearl Account of Singular Causation

Halpern & Pearl (2001, 2005) present an influential account which attempts to derive facts about singular causation from structural determination relations amongst variables, along with a specification of the variable's actual values. Some notation: given a structural equations model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ and a specification of the variable values at a world ω at which $\mathcal{M}$ is correct, $\mathcal{U}_\omega$, we can write that

$$\mathcal{M}, \mathcal{U}_\omega \models \mathscr{A}$$

iff $\mathscr{A}$ is true of the system of equations $\mathcal{M}$, given the assignment $\mathcal{U}_\omega$.[3] Then, the Halpern & Pearl account of singular causation is as follows.[4]

(HP) $C = c$ is a singular cause of $\mathbf{E} = \mathbf{e}$ at world ω, according to $\mathcal{M}$, iff

1) $\mathcal{M}, \mathcal{U}_\omega \models C = c \wedge \mathbf{E} = \mathbf{e}$

2) There is a partition of $\mathcal{U} \cup \mathcal{V}$ into $\mathbf{P}$ and $\mathbf{O}$, such that $C \in \mathbf{P}$, and assignments of values c' and $\mathbf{o}$ such that

[3] Given the values of the exogenous variables, $\mathcal{U}_\omega$, and the structural equations in $\mathcal{E}$, we can work out the values of all the endogenous variables, $\mathcal{V}_\omega$.

[4] This is not exactly the account presented in Halpern & Pearl (2001, 2005). That account allowed arbitrary *vectors* of variable values $\mathbf{C} = \mathbf{C}_\omega$ to be causes, and additionally included a minimality condition requiring that there not be any subvector of $\mathbf{C} = \mathbf{C}_\omega$ which also meets conditions (HP, 1) and (HP, 2). Eiter & Lukasiewicz (2002) proved that, given the 2001 HP account, the minimality condition will always winnow $\mathbf{C} = \mathbf{C}_\omega$ down to a single variable value, so I have decided to sacrifice generality for perspicacity by presupposing that the cause will be a single variable's value and ditching the minimality condition.

Things are slightly complicated by the fact that the account was slightly modified between Halpern & Pearl (2001) and Halpern & Pearl (2005) to include the requirement that (HP, 2, b) hold for any *subvector* $\mathbf{O}^* \subseteq \mathbf{O}$, and Halpern (2008) demonstrates that, with this addition to the account, the result of Eiter & Lukasiewicz, that the minimality condition will always winnow $\mathbf{C} = \mathbf{C}_\omega$ down to a single variable value, no longer holds. Since I will only consider cases in which $\mathbf{O}$ is unitary, there will not be any difference between the 2001 and 2005 account with respect to any of the cases I'm considering (which is why I've chosen the simpler 2001 account), so it's safe for me to neglect the minimality condition and presuppose that the cause will be a single variable value.

a) $\mathcal{M}, \mathcal{U}_\omega \models (C = c' \wedge \mathbf{O} = \mathbf{o}) \mathrel{\Box\!\!\rightarrow} \mathbf{E} \neq \mathbf{e}$
and

b) $\mathcal{M}, \mathcal{U}_\omega \models (C = c \wedge \mathbf{O} = \mathbf{o} \wedge \mathbf{P}^* = \mathbf{P}^*_\omega) \mathrel{\Box\!\!\rightarrow} \mathbf{E} = \mathbf{e}$
for all $\mathbf{P}^* \subseteq \mathbf{P}$.

Condition (HP, 1) simply requires that, as a matter of fact, both $C = c$ and $\mathbf{E} = \mathbf{e}$ actually occur. The interesting part of the account comes in condition (HP, 2). (HP, 2) is something of a counterfactual criterion, stating that $\mathbf{E} = \mathbf{e}$ counterfactually depends upon $C = c$. However, (HP, 2) does not require this counterfactual dependence to obtain given the *actual* assignment of variable values $\mathcal{U}_\omega$. Rather, it merely requires that there be some assignment of variable values (possibly, but not necessarily, the actual assignment) to a vector of variables $\mathbf{O}$ such that, holding those variables fixed at *that* assignment, $\mathbf{E} = \mathbf{e}$ counterfactually depends upon $C = c$. Though the account doesn't specify which variables go in $\mathbf{P}$ and $\mathbf{O}$, it helps to get a feel for the account to think of $\mathbf{P}$ as a vector of the variables laying on some *p*ath between C and the variables appearing in $\mathbf{E}$, and to think of $\mathbf{O}$ as a vector of the *o*ff-path variables. Then, condition (HP, 2, a) requires that, were we to hold the off-path variables fixed at some (possibly non-actual) values, if C were set to some value $c' \neq c$, $\mathbf{E}$ would not be $\mathbf{e}$. Condition (HP, 2, b) is meant to require that it is not our holding the off-path variables fixed which is, all by itself, responsible for the changes in $\mathbf{E}$. The idea is this: if, holding fixed any number of the variables on the path between C and $\mathbf{E}$ at their actual values, changing the values of the off-path variables to $\mathbf{o}$ does not by itself result in a change in $\mathbf{E}$, then these changes in the off-path variables are not entirely responsible for the change in E; therefore, $C = c'$ must be partly responsible for these changes, in the counterfactual scenario where $C = c' \wedge \mathbf{O} = \mathbf{o}$.

In both the 2001 and the 2005 Halpern & Pearl articles—as well as in Halpern (2008) and Halpern & Hitchcock (2010)—this twist in the simple counterfactual condition is motivated by appeal to symmetric overdetermination cases. To use the example that appears in all of those articles, suppose that, given the circumstances, either a lit match or lightning would be sufficient for a forest fire, so that the structural equation $F := M \vee L$ is true (where all these variables are binary and F is 1 iff there is a forest fire, M is 1 iff the match is lit, L is 1 iff there is lightning, and $\circ \vee *$ is the familiar truth function $\max\{\circ, *\}$). In the case where $M = L = 1$, the simple counterfactual test will fail for both $M = 1$ and $L = 1$, so the simple counterfactual test will

$$\mathcal{E}_{16} = \big(W := S \wedge \overline{P}\big)$$

$$\begin{tikzcd} S \arrow[r] & W \\ P \arrow[ur] & \end{tikzcd}$$

Figure 16

claim that neither the match nor the lightning caused the forest fire.[5] Halpern & Pearl (2001, 2005), Halpern (2008), and Halpern & Hitchcock (2010) deem this result unacceptable; they claim that it is a desideratum of an account of singular causation that it say that both $M = 1$ and $L = 1$ individually cause $F = 1$. They diagnose the problem with the standard counterfactual criterion thusly: $L = 1$ serves to *mask* the causal influence of $M = 1$ on $F = 1$. However, we can reveal this influence by considering the counterfactual setting $L = 0, M = 1, F = 1$. In this counterfactual setting of the variables, $F = 1$ *will* counterfactually depend upon $M = 1$; so, on the (HP) account, given the actual variable settings, $M = 1$ causes $F = 1$.

Unfortunately, (HP) is all too easily satisfied. Just because $\mathbf{E} = \mathbf{e}$ counterfactually depends upon $C = c$ in some *counterfactual* world, this does not mean that $C = c$ causes $\mathbf{E} = \mathbf{e}$ in the *actual* world. Here's a case which demonstrates the problem: Suzy stands in front of a window with a rock in hand, and there is a partition raised which separates Suzy from the window. Suzy does not throw her rock; and the window does not shatter. Were Suzy to throw the rock with the partition lowered, the window would have shattered. However, were Suzy to throw the rock with the partition raised, then the window would not have shattered. Then, given a counterfactual account of the truth conditions of structural equations models—which Halpern & Pearl (2005) appear to accept[6]—the system of structural equations $\mathcal{E}_{16}$, shown in figure 16, should be correct. (In figure 16, W, S, and P are binary variables with the natural interpretation, $\circ \wedge *$ is the truth function $\min\{\circ, *\}$, and $\overline{*}$ is the truth function $1 - *$.)

[5] Although, note, if we allow disjunctions of variable values to be causes, it *will* get the result that the disjunction $M = 1 \vee L = 1$ caused the forest fire. (This is the approach to symmetric overdetermination that I endorse in §5.3.9.)

[6] See Halpern & Pearl 2005, p. 847: "An equation such as $Y := \phi_Y(X)$ should be thought of as saying that...if X were set to x by some means (not specified in the model), then Y would take on the value $\phi_Y(x)$, as dictated by ϕ_Y." (Here, I have changed Halpern & Pearl's notation to match my own.)

And the actual variable values are

$$P = 1$$
$$S = 0$$
$$W = 0$$

Now, (HP) ends up entailing that $P = 1$ causes $W = 0$. That's because both $P = 1$ and $W = 0$ are true in the actual context, so (HP, 1) is satisfied. And we can partition the variables into $\mathbf{P} = (P, W)$ and $\mathbf{O} = (S)$, and there are assignments $P = 0$ and $S = 1$, such that

$$(P = 0 \land S = 1) \mathrel{\Box\!\!\rightarrow} W = 1$$

and

$$(P = 1 \land S = 1) \mathrel{\Box\!\!\rightarrow} W = 0$$

So (HP, 2) is satisfied.[7] So, according to (HP), the partition being up kept the window from shattering.[8] But this is crazy. Suzy didn't even throw her rock. It's true that, if Suzy *had* thrown, then the partition being up *would have* kept the window from shattering. But, as it is, the partition simply didn't have anything to do with the window's not shattering.

Hall (2007) presents a deeper problem. It is not an objection to the particulars of (HP), but rather an objection to the very project of providing an account of singular causation in terms of causal models and variable values in the first place. The problem Hall poses is that there are systems of structural equations which are isomorphic to one another, and in which all the variables take on the same values, but which nevertheless differ in their singular causal structure.

First, Hall calls our attention to the *neuron diagram* (in the style of Lewis, 1986a) displayed in figure 17. Here's how to read the diagram in figure 17: for every time t listed at the bottom, the neurons above it can either fire or not fire at that time. If a neuron actually fires at its designated time, then it is colored gray. The arrows represent *stimulatory* connections between neurons: if the neuron at the tail of the arrow fires at its designated time, then, *ceteris paribus*, the neuron at the head will fire at its designated time. The lines with circles at their heads give the reason for the

7 (HP, 2, b) will also require the truth of the counterfactual '$(P = 1 \land S = 1 \land W = 0) \mathrel{\Box\!\!\rightarrow} W = 0$'.

8 Slightly more carefully: the partition being up caused the window to not shatter. (We might worry that the causative 'kept from shattering' says more than 'caused to not shatter'.)

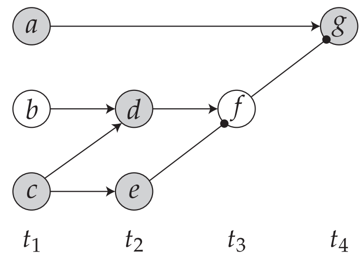


Figure 17

$$\mathcal{E}_{18} = \begin{pmatrix} g & := & a \wedge \overline{f} \\ f & := & d \wedge \overline{e} \\ d & := & b \vee c \\ e & := & c \end{pmatrix}$$

$$\begin{tikzcd} b \arrow[r] & d \arrow[r] & f \arrow[r] & g \\ & c \arrow[u] \arrow[r] & e \arrow[u] & a \arrow[u] \end{tikzcd}$$

Figure 18

ceteris paribus clause in the previous sentence; they represent *inhibitory* connections between neurons: if the neuron at the base of the line fires, then the neuron at the head of the line *won't* fire. So, for instance, if f fires at t_3, then g will not fire at t_4 (whether a fires or not).

Hall points out that we can model this neuron diagram with the set of structural equations $\mathcal{E}_{18}$, shown in figure 18. In figure 18, each of the variables take the value 1 if their associated neurons fire and 0 if they don't. $\mathcal{E}_{18}$ tells us that g will fire iff a fires and f does not fire; that f will fire iff d fires and e does not; that d will fire iff either b or c fires; and that e will fire iff c does.

Hall then asks us to consider the neuron diagram shown in figure 19. There, the double-circle around G indicates that it is a dull neuron — it requires *two* stimulatory signals in order to fire. So, it will fire iff both A and F fire. The reason for drawing our attention to this neuron diagram is that it can be modeled with a set of structural equations, $\mathcal{E}_{20}$, shown in figure 20, which is *isomorphic* to the first. In figure 20, following the same conventions as before, $F^* \stackrel{\text{def}}{=} \overline{F}$, $D^* \stackrel{\text{def}}{=} \overline{D}$, and $B^* \stackrel{\text{def}}{=} \overline{B}$. Moreover, not only are the two structural equations isomorphic in this case, but the actual values of the corresponding variables are the same:

$$\begin{array}{lll} a = A = 1 & d = D^* = 1 & g = G = 1 \\ b = B^* = 0 & e = E = 1 & \\ c = C = 1 & f = F^* = 0 & \end{array}$$

Thus, if we are to extract singular causal information from $\mathcal{M}$ and $\mathcal{U}_\omega$ alone, then

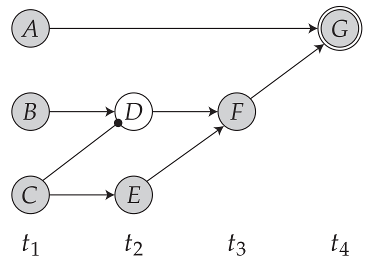


Figure 19

$$\mathcal{E}_{20} = \begin{pmatrix} G & := & A \wedge \overline{F^*} \\ F^* & := & D^* \wedge \overline{E} \\ D^* & := & B^* \vee C \\ E & := & C \end{pmatrix}$$

$$\begin{tikzcd}
B^* \arrow[r] & D^* \arrow[r] & F^* \arrow[r] & G \\
 & C \arrow[u] \arrow[r] & E \arrow[u] & A \arrow[u]
\end{tikzcd}$$

Figure 20

the singular causal structure of the two neuron diagrams should be the same. But the singular causal structure of the two neuron diagrams *aren't* the same. In the first, the firing of c doesn't cause the firing of g; while, in the second, the firing of C *does* cause the firing of G.

So the causal model and the actual values of the variables alone don't give us enough information to tell which events cause which other events. Hall (2007), Hitchcock (2007), and Halpern (2008) deal with this problem by throwing information about a variable's *default*, *normal*, or *typical* values into the mix as well. By way of explanation: some authors[9] have thought that either judgements of *default* versus *deviation*, or else comparative notions of normality or typicality, are central to our understanding of causation—*e.g.*, Maudlin (2004) suggests that all causal claims presuppose laws governing how something will behave if it is left alone, its *default* condition, along with laws detailing conditions under which the system will deviate from its default state, and prescribing *how* it will so deviate. And *not* firing is the default behavior of a neuron, whereas *firing* is a deviation therefrom. Note that in the neuron diagram in figure 17, d does fire, while in the neuron diagram in figure 19, D doesn't fire. If we're persuaded that the default/deviation distinction, or some more comparative notion of normality or typicality, is central to causation, then we'll think that these are two very different kinds of events. However, both of these events just

[9] Halpern (2008) points in particular to the psychological work of Kahneman & Miller (1986). Hitchcock & Knobe (2009) contend that there is an important relationship between judgments of singular causation and judgments of normality or typicality, and situate their account in a framework of structural equations models.

get represented in the structural equations model by the associated variable taking the value 1. The model thus ignores default, normality, or typicality information. And the obvious way of remedying this defect is to just include this information in our models.

Halpern (2008) achieves this by enriching structural equations models with a *ranking function* r which maps assignments of variable values to the positive integers $\mathbb{Z}^+$—variable assignments mapped to lesser integers are more normal or typical than those mapped to greater integers. Then, condition (HP, 2) is changed to read:

2) There is a partition of $\mathcal{U} \cup \mathcal{V}$, $(\mathbf{P}, \mathbf{O})$, such that $C \in \mathbf{P}$, and there is some assignment $\mathcal{U} \cup \mathcal{V} = \mathbf{u} \cup \mathbf{v}$ such that $r(\mathbf{u} \cup \mathbf{v}) \leq r(\mathcal{U}_\omega \cup \mathcal{V}_\omega)$, and according to which $C = c'$ and $\mathbf{O} = \mathbf{o}$, and

   a) $\mathcal{M}, \mathcal{U}_\omega \models C = c' \wedge \mathbf{O} = \mathbf{o} \mathbin{\Box\!\!\rightarrow} \mathbf{E} \neq \mathbf{e}$
   and
   b) $\mathcal{M}, \mathcal{U}_\omega \models C = c \wedge \mathbf{O} = \mathbf{o} \wedge \mathbf{P}^* = \mathbf{P}^*_\omega \mathbin{\Box\!\!\rightarrow} \mathbf{E} = \mathbf{e}$
   for all $\mathbf{P}^* \subseteq \mathbf{P}$.

In other words, the counterfactual scenario in which you test for counterfactual dependence must be at least as *normal* or *typical* as the actual scenario.

To get a feel for this extra requirement, think back to the counterexample I presented to the HP account on page 109—in which Suzy did not throw her rock at a window which was protected by a raised partition. The souped-up version of the (HP) account does not get the result that the partition being raised caused the window to not shatter, so long as we suppose that the assignment $S = 1$ is less normal than the actual assignment $S = 0$; so long, that is, as we suppose that not throwing is Suzy's default state, and throwing is a deviation therefrom.

4.3 Problems with Normality

Unfortunately, I don't think that the souped-up (HP) account is going to work either. Suppose there is a live grenade sitting with Suzy and the window. If the grenade explodes where it sits, then the shock waves will shatter the window, unless a partition protecting the window is raised. If Suzy throws the grenade away, then the window will not shatter. So, the window will shatter iff Suzy *doesn't* throw and the partition is down. So the structural equations model $\mathcal{E}_{21}$, shown in figure 21, will be correct.

$$\mathcal{E}_{21} = (W := \overline{S} \wedge \overline{P})$$

$$\begin{tikzcd} S \arrow[r] & W \\ & P \arrow[u] \end{tikzcd}$$

Figure 21

(There, W, S, and P are binary variables with the natural interpretation.) Suppose that the partition is up and Suzy throws, so that the actual assignment of values are

$$S = 1$$
$$P = 1$$
$$W = 0$$

Now, it is actually true that $P = 1$ and $W = 0$. And I can partition the variables into $\mathbf{P} = (P, W)$ and $\mathbf{O} = (S)$, and it is true that

$$(P = 0 \wedge S = 0) \mathrel{\Box\!\!\rightarrow} W = 1$$

and that

$$(P = 1 \wedge S = 0) \mathrel{\Box\!\!\rightarrow} W = 0.$$ [10]

And, given that not throwing is a default and throwing is a deviation therefrom, the assignment $S = 0$ should be at least as normal as $S = 1$.

So the souped-up (HP) definition rules that the partition being up caused the window to not shatter, even though Suzy threw the grenade away. This is a case of *preemptive prevention*.[11] These cases are interesting, since it is not clear whether we ought to count Suzy's throwing of the grenade as a cause of the window's failure to shatter. We can reason that, either Suzy's throwing it or the partition's being up saved the window, and it wasn't the partition's being up that did it, so it must have been Suzy's throwing the grenade that did it. Or we can reason that, since the partition would have protected the window even if she didn't throw it, the throw didn't actually save the window. Which of these reasonings we should favor is controversial. What ought *not* be controversial is that, given that Suzy actually threw the grenade away, the partition's being up didn't save the window. The souped-up (HP) account says otherwise. So much the worse for the souped-up (HP) account.

[10] The revised (HP, 2, b) will also require the truth of the counterfactual '$(P = 1 \wedge S = 0 \wedge W = 0) \mathrel{\Box\!\!\rightarrow} W = 0$'.

[11] See McDermott (1995, p. 525). These kinds of cases are discussed in greater detail in Collins (2004).

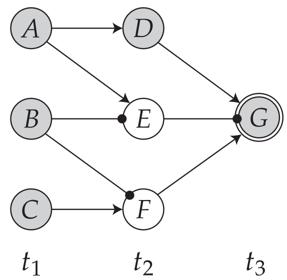


Figure 22

$$\mathcal{E}_{23} = \begin{pmatrix} D := A \\ E := A \wedge \overline{B} \\ F := C \wedge \overline{B} \\ G := (\overline{E} \wedge (D \vee F)) \vee (D \wedge F) \end{pmatrix}$$

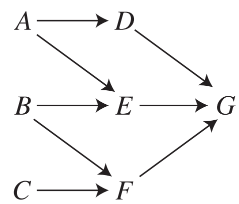


Figure 23

A defender of the souped-up (HP) account might try to say that, while throwing *rocks* is atypical, throwing *grenades* is typical. Such *ad hoc* normality claims are a clever strategy for avoiding counterexamples—too clever by half, if you ask me—but the mere fact that such maneuvers are necessary should lessen our confidence that the account is on the right track. In any case, we can produce a counterexample to the souped-up (HP) account using simple neuron diagrams, and everybody seems to agree about how the default-deviation distinction is to be applied to neuron diagrams—to remain dormant is default; to fire is deviant. Consider, then, the neuron diagram shown in figure 22. In this neuron diagram, the inhibitory signal from E cancels out any *one* stimulatory signal to G. So G will fire just in case E *doesn't* fire and either D or F fires, or else E *does* fire and *both* D and F fire. This neuron diagram can be represented with the system of structural equations $\mathcal{E}_{23}$, shown in figure 23. There, A, B, C, D, E, F, and G are all binary variables which take the value 0 if their associated neurons do not fire at their associated times and take the value 1 if they do. Now, we can partition the variables into $\mathbf{P} = (A, B, D, E, F, G)$ and $\mathbf{O} = (C)$, and it is true that

$$(B = 0 \wedge C = 0) \mathrel{\Box\!\!\rightarrow} G = 0$$

and it is also true that

$$(B = 1 \wedge C = 0 \wedge \mathbf{P}^* = \mathbf{P}^*_{@}) \mathrel{\Box\!\!\rightarrow} G = 1$$

for all $\mathbf{P}^* \subseteq \mathbf{P}$—since, in the counterfactual scenario in which B fires and C does not, there will be no difference between the actual and the counterfactual values of any of the variables in $\mathbf{P}$. In other words: if we have C *not fire*—which, everybody seems to agree, is more normal or typical than C firing—then whether G fires counterfactually depends upon whether B fires; so the normality-based HP account of causation entails that B's firing caused G to fire. However, given that A and C both actually fired, B's firing was *not* a cause of G's firing. Even if B's firing caused G's firing to be caused by D alone, rather than jointly caused by D and F, B's firing didn't cause G to fire.

4.4 More Underdetermination Problems

It gets worse. Not only does Halpern (2008)'s normality-based account fail to work, but it can be shown that *no* normality-based approach can be made to work, given a counterfactual understanding of structural equations models. Further, no normality-based approach can be made to work, so long as our account of the correctness conditions for causal models satisfies three independently plausible conditions. Both the causal counterfactual understanding and the nomic sufficiency understanding of structural equations models sketched in the previous chapter are committed to these three principles. And if we accept these three principles, then there will be pairs of isomorphic structural equations models, each of which are correct, whose variables take on all the same actual values, and where each of those values have the same normality ranking, but which nevertheless differ with respect to which variable values singularly caused which other variable values.

To begin with, I will assume that, given a neuron diagram, a system of structural equations which consists entirely of binary variables for each neuron, together with equations specifying the conditions under which the neurons will fire, will constitute a *correct* structural equations model of that neuron diagram. Call a structural equations model like this the *canonical representation* of a neuron diagram. The first principle I will require, then, is that the canonical representation of a neuron diagram is correct.

The Canonical Representation of Neuron Diagrams is Correct. *A neuron diagram is correctly modeled by a variable for each neuron which takes the value* 0 *if the neuron doesn't fire and takes the value* 1 *if it does, and a system of structural equations specifying which other neurons each neuron immediately depends upon and how it so depends.*

This principle follows from the counterfactual understanding of structural equations models, since all of the counterfactuals associated with such a structural equations model will be true. It also follows from the nomic sufficiency account sketched in the previous chapter.

A quick word on neuron diagrams: Lewis (1986a) introduced neuron diagrams as a means for *representing* some part of the actual world's causal structure. However, we need not understand neuron diagrams as representational tools. We can simply understand them as completely describing a self-contained system which follows certain very simple laws. There are possible worlds which consist entirely of neurons connected in the ways specified by the diagram and obeying the law 'fire only if you receive a stimulatory signal, unless you also receive an inhibitory signal'. Such possible worlds make good test cases for thinking about the world's causal structure, since often enough, it is very clear which neural firings caused which other neural firings, and which did not. In their simplicity, these neuron diagrams provide a minimal bar which an account of causation ought to meet. Where it is clear which neural firings caused which other neural firings, it is a desideratum of an account of causation that it get these cases right. If the account can't be made to work in the simple and sterile world of neuron diagrams, then it can't be expected to work in the rich and messy actual world.

We need not include *every* feature of a system in our structural equation model. It is fine for a model to take some features of the world for granted. When we model Suzy and Billy throwing their rocks at the window, we need not include a variable for the moon's gravitational pull. Since the moon's gravitational pull is constant, our model is no worse for not including this variable—even though, at some counterfactual values, this variable *would* make a difference to whether the window shatters. The same ought to be true of neuron diagrams. We need not include a variable for *every* neuron in the diagram. So long as the counterfactuals entailed by the system of equations are all true, it doesn't matter whether *every* neuron is explicitly represented. For instance, in the neuron diagram displayed in figure 17 (reproduced above for convenience), we need not include a variable for the neuron b in the diagram. Given that b didn't actually fire, this neuron diagram would also be adequately described by the system of structural equations $\mathcal{E}_{24}$ shown in figure 24. Of course, if we wanted to reason about what would happen if we were to intervene upon neuron b, we would have to include such a variable in our model. But that doesn't mean that the pared-down structural equations model can't also be correct, and that we can't appropriately use it to determine the causal relationships between, *e.g.*, the firing of c and the firing of e. So it seems that we should accept the following principle.

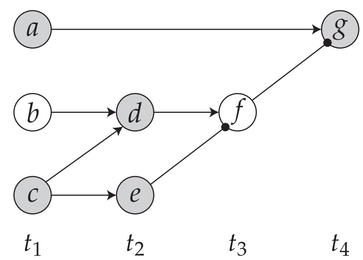


Figure 17

$$\mathcal{E}_{24} = \begin{pmatrix} g & := a \wedge \overline{f} \\ f & := d \wedge \overline{e} \\ d & := c \\ e & := c \end{pmatrix}$$

$$\begin{tikzcd} a \arrow[r] & g \\ d \arrow[r] & f \arrow[u] \\ c \arrow[u] \arrow[r] & e \arrow[u] \end{tikzcd}$$

Figure 24

Removing Exogenous Variables Preserves Correctness. *If a structural equations model* $< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ *is correct, then any structural equations model attainable by removing an exogenous variable* $U \in \mathcal{U}$ *is also correct.*

The way to remove an exogenous variable is just to exchange the variable U for its actual value u in all structural equations in which U appears. Again, this principle follows from the counterfactual understanding of structural equations models, since the counterfactuals entailed by the pared-down model is just a subset of the counterfactuals entailed by the original model. If the original model was correct, then the pared-down model ought to be correct as well. The same goes for the nomic sufficiency account from chapter 3.

Similarly, we need not include every intermediate step between two events in our structural equations model. So long as there is the right kind of counterfactual covariation between values of U and V, it shouldn't matter which other variables sit intermediate between U and V. For instance, we need not include a variable for the firing of d in the model shown in figure 17. Since d takes on the same value as c, we can simply exchange c for d in f's structural equation and get the system of equations $\mathcal{E}_{25}$ shown in figure 25. Similarly, since f's value is just a function of c and e, we can reduce the system of equations further to get the system of equations shown in figure 26. The counterfactuals entailed by this system of structural equations are all true; and there are not any extra counterfactual relationships between the variables a, c, e,

$$\mathcal{E}_{25} = \begin{pmatrix} g & := a \wedge \overline{f} \\ f & := c \wedge \overline{e} \\ e & := c \end{pmatrix}$$

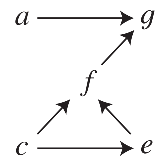


Figure 25

$$\mathcal{E}_{26} = \begin{pmatrix} g := a \wedge (\overline{c} \vee e) \\ e := c \end{pmatrix}$$

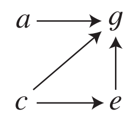


Figure 26

and g which the structural equations model fails to capture. If the original variable set was an appropriate variable set, then the new variable set should be appropriate as well, since the new variable set is just a subset of the original. So the structural equations model $\mathcal{E}_{26}$ ought to be correct.

In general, it seems that we should accept the principle that functional composition of the equations in $\mathcal{E}$ preserves the correctness of a structural equations model.

Functional Composition Preserves Correctness. *If a structural equations model $< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ is correct, then any structural equations model attainable by taking the functional composition of equations in $\mathcal{E}$ and removing the intermediate variables and their structural equations is correct as well (so long as the composed function is not constant).*

This principle also follows from the counterfactual understanding of structural equations models; as well as the nomic sufficiency account. (It is important, by the way, that we remove the intermediate variables and their equations from the model, since otherwise, the resulting system of equations may entail false counterfactuals. For instance, if we replace g's equation with $g := a \wedge (\overline{c} \vee e)$ but retain the equation $f := c \wedge \overline{e}$, then the resulting system of equations will entail that, if we were to intervene so as to set f to 1 and c to 0, then g would still fire, which is false.)

The problem is that, if we accept each of these three principles, then we ought to think that the system of equations $\mathcal{E}_{26}$ correctly describes *both* the neuron diagram from figure 17 *and* the neuron diagram shown in figure 27. Since, given these three principles, we can start with the canonical representation of the neuron diagram on display in figure 27, shown in figure 28. And, by removing the exogenous variable B, get the system of equations $\mathcal{E}_{29}$, shown in figure 29. Then, by functional composition,

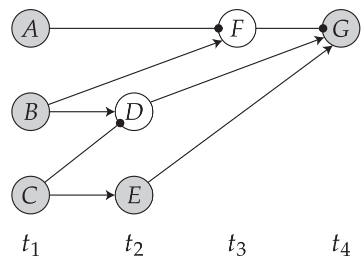


Figure 27

$$\mathcal{E}_{28} = \begin{pmatrix} G := \overline{F} \wedge (D \vee E) \\ F := \overline{A} \wedge B \\ D := B \wedge \overline{C} \\ E := C \end{pmatrix}$$

$$\begin{tikzcd} A \arrow[r] & F \arrow[r] & G \\ B \arrow[ur] \arrow[r] & D \arrow[ur] & \\ C \arrow[ur] \arrow[rr] & & E \arrow[uu] \end{tikzcd}$$

Figure 28

we can remove the variables F and D, getting the system $\mathcal{E}_{30}$, shown in figure 30, which is isomorphic to the system $\mathcal{E}_{26}$.

Not only are the structural equations the same, but the *values* of the variables are the same:

$$a = c = e = g = A = C = E = G = 1$$

And, in each case, the assignment 1 to a variable corresponds to that variable firing at its designated time. So there's no difference between the two models with respect to the normality of their variable assignments.

However, in the neuron diagram on display in figure 17, the firing of c *didn't* cause the firing of g; whereas, in the neuron diagram on display in figure 27, the firing of C *did* cause the firing of G.

The lesson: if we accept the three principles about the correctness of structural equations models laid down above, then we must think that a correct structural equations model and a normality ranking of variable values underdetermines the singular causal relations between the values of the variables appearing in that model. Equivalently: if we want to be able to derive relations of singular causation from a correct structural equations model, the standards of correctness must be strong enough to invalidate one of these three principles.

$$\mathcal{E}_{29} = \begin{pmatrix} G := \overline{F} \wedge (D \vee E) \\ F := \overline{A} \\ D := \overline{C} \\ E := C \end{pmatrix}$$

$$\begin{tikzcd} A \arrow[r] & F \arrow[r] & G \\ & D \arrow[ur] & \\ C \arrow[ur] \arrow[rr] & & E \arrow[uu] \end{tikzcd}$$

Figure 29

$$\mathcal{E}_{30} = \begin{pmatrix} G := A \wedge (\overline{C} \vee E) \\ E := C \end{pmatrix}$$

$$\begin{tikzcd} A \arrow[r] & G \\ C \arrow[ur] \arrow[r] & E \arrow[u] \end{tikzcd}$$

Figure 30

4.5 Objections and Suggestions

Rejecting recalcitrant data is always an option. We could, then, simply deny that the firing of C caused the firing of G in the neuron diagram in figure 27. Alternatively, we could deny that the firing of c *didn't* cause the firing of g in the neuron diagram in figure 17. To do this, however, is only to solve one instance of a more general problem. Other neuron diagrams can be constructed which, together with the three principles of the previous section, give rise to a conflict between our theory and our characteristic causal judgments. If we take the strategy of rejecting recalcitrant data, there may be quite a lot of recalcitrant data to reject. I take this to be a significant theoretical cost; and an option of, if not last resort, then at least *late* resort.

Another potential response to these kinds of worries shows up in Halpern & Hitchcock (2010), where we are told that

> the structural equations can be viewed as describing objective [counterfactual] features of the world. However...the choice of variables and their values can also have a significant impact on causality. Moreover, these choices are, to some extent, subjective. This, in turn, means that judgments of [singular] causation are subjective.[12]

These comments about subjectivism could be taken in at least two ways. Firstly, we might suspect that Halpern & Hitchcock are endorsing a kind of contextualism about claims of singular causation; so that two different contexts could raise to salience two different structural determination relations, and the salience of those structural

[12] Halpern & Hitchcock (2010, §1)

determination relations could make a difference to the proposition expressed by a sentence of the form '*c* caused *e*'. We might then think that two correct structural equations models could differ in their verdicts about which variable values are causally related by representing, or failing to represent, these structural determination relations.[13] If we think this, we might be inclined to reject the counterexample of §4.4 by claiming either that there are contexts within which the sentence '*C*'s firing didn't cause *G*'s firing' is true, or that there are contexts within which the sentence '*c*'s firing caused *g*'s firing' is true, and that the verdicts of the pared-down structural equations models $\mathcal{E}_{26}$ and $\mathcal{E}_{30}$ align with these contexts.

If there's a context within which it's true to say '*c*'s firing caused *g*'s firing', it's a context in which we hold fixed (for whatever reason) the fact that *d* fired. In that case, *d* would represent a threat to *g*'s firing which *c* diffused by causing *e* to fire. But we ought not think that the pared-down structural equations model $\mathcal{E}_{26}$ puts us in *this* context, since it doesn't even include a variable for *d*. It therefore doesn't allow us to consider the possibility in which *d* fires even though *c* doesn't. And if there's a context in which it's true to say '*C*'s firing *didn't* cause *G*'s firing', it's a context in which we are thinking about the lack of counterfactual dependence between *C*'s firing and *G*'s firing—since, had *C* not fired, *B* would have made *G* fire. However, there is a lack of counterfactual dependence between *C*'s firing and *G*'s firing in *both* the original structural equations model *and* in the pared-down system of equations $\mathcal{E}_{30}$. Moreover, only the original model allows us to explicitly consider the possible influence of *B*'s firing. So I'm dubious that contextualism is going to give us the leverage to diffuse the counterexample of §4.4.

Alternatively, we might find in Halpern & Hitchcock's comments a deeper kind of subjectivism, according to which there either is no fact of the matter, contextually-dependent or not, about which variable values singularly cause which other variable values, or else these facts vary from person to person, depending upon which structural equations model they use to model the situation. In either case, if we are to be in the business of giving an *account* of anything at all, then we must be in the business of providing an account of causal *judgments*. And the claim must be that causal judgments vary from person to person depending upon which structural equations model they use to model the situation.[14] Speaking for myself, I do not find my causal

[13] See Menzies (2007) for a development of this kind of contextualism about singular causation.

[14] In that case, talk of 'modeling a situation' must be taken to be metaphorical; as the overwhelming majority of causal judgments precede the mathematical formalism of structural equations models.

judgments shifting when I model either of the neuron diagrams with the system of equations $\mathcal{E}_{26}$ or $\mathcal{E}_{30}$; nor when I draw my attention only to the neurons A, C, E, and G (or a, c, e, and g) and the counterfactual relationships between them (nor do the judgments shift for a smattering of philosophers and non-philosophers to whom I have presented the case). If this is right, then subjectivism is just another instance of the strategy 'deny the recalcitrant data', bundled up with extra metaphysical baggage. As I said, this should be a strategy of late resort. (Perhaps there are other reasons to favor this brand of subjectivism, but it does not receive motivation from, nor provide a convincing response to, the counterexample of §4.4.)

(As a matter of exegesis, it seems to me that another, perhaps more charitable reading of Halpern & Hitchcock is suggested by Halpern (ms)'s comment that "As [Halpern & Hitchcock (2010)] pointed out, constructing a good model is *still* more of an art than a science."[15] The "still" here suggests that the fact that constructing a good model is more of an art than a science—and the concomitant subjectivism—are understood as an ideally temporary state of affairs; the result of the research program being still in its infancy. This reading also jibes better with some of the comments I'll discuss below.)

There is another kind of response to underdetermination problems like these that shows up in Halpern & Pearl (2005) and Halpern & Hitchcock (2010). In each of these articles, the authors consider the well-worn preemption case from Lewis (1986a) in which Suzy and Billy both throw their rocks at the window, but Suzy's rock arrives first. As a first pass, it appears that we can properly model this case with three binary variables, one which is 1 iff Suzy throws her rock and is 0 otherwise, one which is 1 iff Billy throws his rock and is 0 otherwise, and one which is 1 iff the window shatters and is 0 otherwise. With these variables, all of the causal counterfactuals entailed by the system of structural equations

$$\mathcal{E} = (W := S \vee B)$$

are true. However, this equation is perfectly symmetric with respect to S and B; and if $S = B = 1$, then there is nothing to distinguish $S = 1$ from $B = 1$. But, in the standard story, it was *Suzy's* throw, and not Billy's, which caused the window to shatter, since Suzy's rock was the one that actually *hit* the window. The window was already shattered by the time Billy's rock arrived. It could also be the case, consistent

[15] Halpern (ms, §6, emphasis added)

with the structural equation $W := S \vee B$, that $S = B = 1$, yet *Billy's* rock was the one that shattered the window. The model, therefore, underdetermines the singular causal structural of the variables S, B, and W. This is another flavor of the kind of underdetermination worry that I've been discussing for the past twenty-odd pages. Halpern & Pearl (2005) and Halpern & Hitchcock (2010) both consider this causal model, and both deem it inadequate. However, neither have much to say about *why* the model is inadequate. Both simply say:

> The trouble with this model is that it cannot distinguish the case where both rocks hit the [window] simultaneously...from the case where Suzy's rock hits first. (Halpern & Pearl 2005, p. 861, Halpern & Hitchcock 2010, § 3.2, p. 11)

Both are quick to point out that there are *other* models of the Suzy and Billy case which can be used to ground an asymmetry between Suzy's and Billy's throw.

These remarks suggest a simplistic strategy for dealing with any underdetermination problem (which I do not attribute to these authors): just reject the model within which singular causation is underdetermined, and suggest an alternate model. Perhaps we also throw in some very specific remarks about what was wrong with the original model that doesn't clearly generalize to any other models. There is something frustratingly evasive about this strategy. Halpern & Pearl (2005) and Halpern & Hitchcock (2010) are, of course, clearly *right* that what's wrong with the structural equation $W := S \vee B$ is that it doesn't allow us to distinguish the case where both rocks hit simultaneously from the case where one of the rocks hits first. But this remark doesn't tell us anything about what more than the correctness of the causal counterfactuals is required for a structural equations model to be adequate. It may, of course, be a good strategy for practicing researchers: if a model yields results known to be false, find a new model. But it does nothing to address the question of what it is for one of these models to be correct in the first place. And, presumably, if we're in the business of providing an account of singular causation, we will want to know how to answer this question.

We may well also wonder why, if this is an acceptable strategy to adopt in response to the the problematic structural equation $W := S \vee B$, isn't it *also* an appropriate strategy to adopt in response to Hall (2007)'s counterexample from §4.2? It is possible, after all, to provide a *new* variable set and a *new* system of structural equations which entail only true counterfactuals and which will join with the (HP) account to deliver the verdict that c's firing didn't cause g's firing in the neuron diagram described in figure 17. (The (HP) account already gets the result that C's firing caused G's firing in the neuron diagram described in figure 19, since, were B to not fire, G's firing would counterfactually depend upon C's firing.) To do this, we just have to remove the variables a, b, and d from the model to get the system of equations

$$\mathcal{E} = \begin{pmatrix} g & := \overline{f} \\ f & := c \wedge \overline{e} \\ e & := c \end{pmatrix}$$

All of the counterfactuals entailed by this model are correct, and, in this model, there is no way of partitioning the variables so as to satisfy (HP, 2).[16]

This points to a deeper problem with this strategy for dealing with underdetermination worries. If we play the game this way—any counterexample is simply evidence that a different structural equations model is called for—then an account of singular causation will be *immune* from counterexamples. Any potential counterexample simply provides evidence that the structural equations model needs to be changed in some way. If we play by these rules, the account of singular causation never has to stick its neck out; it never has to make any predictions at all.

Here's another suggestion. Up until now, I've been running with the idea that, if you have a correct structural equations model $\mathcal{M}$, then all the singular causal relations between the values of the variables in $\mathcal{U} \cup \mathcal{V}$ can be extracted from $\mathcal{M}$ and $\mathcal{U}_\omega$, and perhaps a normality ranking function r. That is, I've been accepting the general approach of Halpern & Pearl (2001, 2005), which is to define singular causation

[16] Halpern (ms) shows that enriching a model with extra *endogenous* variables in a conservative way (in a way such that all the counterfactuals of the original model are still true) can never turn a non-cause into a cause. This case shows that this result cannot be generalized to cases in which we enrich the model with additional *exogenous* variables in a conservative way (a way such that all the counterfactuals of the original model are still true). Since, in the canonical representation of the neuron diagram in figure 17, (HP) counts $c = 1$ as a cause of $g = 1$ (in the counterfactual setting in which $b = 1$, whether $g = 1$ depends upon whether $c = 1$). And this is a conservative extension of the impoverished structural equations model in the body immediately above. However, (HP) does not rule $c = 1$ a cause of $g = 1$ in the impoverished model. So, this is a case in which enriching the model with extra variables changes a non-cause to a cause, according to (HP). Moreover, the present case shows, I think, that this latter notion (enriching by including extra variables, be they endogenous or exogenous) must surely be the notion that we ought to care about when it comes to ferreting out singular causal relations. If start with the impoverished model in the body immediately above, and we want to understand the causal structure of the neuron diagram shown in figure 17, we had better enrich it with additional *exogenous* variables.

only *relative to* a particular structural equations model. However, there are other approaches to accounting for singular causation which insist that there's an important relationship between singular causation and structural determination without claiming that singular causation is only defined relative to a particular structural equations model. For instance, we could account for singular causation by quantifying over *all* correct structural equations models. We could, for instance, claim that $C = c$ is a cause of $\mathbf{E} = \mathbf{e}$ iff there is *some* correct structural equations model within which our account yields the verdict that $C = c$ is a cause of $\mathbf{E} = \mathbf{e}$.[17] This doesn't rule out that there might be *other* correct structural equations models within which our account fails to yield this verdict. If we think about the relationship between causation and structural equations models this way, then we will not be as bothered by false negatives (correct structural equations models within which our account tells us that $C = c$ *doesn't* cause $\mathbf{E} = \mathbf{e}$ even though $C = c$ *does* cause $\mathbf{E} = \mathbf{e}$) as we are by false positives (correct structural equations models within which our account tells us that $C = c$ causes $\mathbf{E} = \mathbf{e}$ even though $C = c$ *doesn't* cause $\mathbf{E} = \mathbf{e}$). For a false negative would still leave open the possibility that some *other* structural equations model gets the judgment that $C = c$ *does* cause $\mathbf{E} = \mathbf{e}$. And notice that the pair of isomorphic structural equations models at the end of §4.4 do not provide a counterexample to *this* way of using structural equations to understand singular causation, since an account of singular causation could rule that $C = 1$ *isn't* a singular cause of $G = 1$ in the structural equations model $\mathcal{E}_{30}$. This would be a false negative, but we could simply look to the more detailed structural equations model to get the verdict that $C = 1$ *is* a singular cause of $G = 1$. Such an account would still stick its neck out. It would open itself up to the risk of false positives within particular structural equations models, as well as the risk of false negatives for which no correct structural equations model can be found which joins with the account to yield the verdict that $C = c$ causes $\mathbf{E} = \mathbf{e}$.[18]

[17] This approach is implicitly adopted in Hitchcock (2001).

[18] Blanchard & Schaffer (forthcoming) criticize this kind of approach on the grounds that there could be one correct structural equations model within which $C = c$ caused $\mathbf{E} = \mathbf{e}$ and $D = d$ *didn't* cause $\mathbf{E} = \mathbf{e}$, and another correct model within which $C = c$ didn't cause $\mathbf{E} = \mathbf{e}$, and $D = d$ did cause $\mathbf{E} = \mathbf{e}$, yet *no* model within which both $C = c$ and $D = d$ caused $\mathbf{E} = \mathbf{e}$. They worry that the above suggestion would yield the result that both $C = c$ and $D = d$ caused $\mathbf{E} = \mathbf{e}$, even though there is no correct structural equation model within which both $C = c$ and $D = d$ caused $\mathbf{E} = \mathbf{e}$. Whether we should be worried by this possibility depends, I believe, upon whether there are possible situations like this. If such a situation were specified, it may constitute a counterexample to the approach; however, we ought not be concerned by the mere fact that out account allows the

While I don't have any objections to this strategy, I am not hopeful that it will solve all of our problems. The reason is that I suspect that singular causation is deeply connected to counterfactual dependence within a correct structural equations model, and I suspect that causation is, in some special circumstances, transitively closed—at times, we can trace out a chain of causal relations and conclude that the first event in the chain caused the last event in the chain, even though the last event doesn't counterfactually depend upon the first (see §5.3.5). So I believe that, given the impoverished structural equations model on display in figure 26, the correct account of singular causation will rule that $c = 1$ causes $e = 1$ (since, in that structural equations model, $e = 1$ counterfactually depends upon $c = 1$), and $e = 1$ will cause $g = 1$ (since, in that structural equations model, $g = 1$ counterfactually depends upon $e = 1$). So $c = 1$ will cause $g = 1$—a false positive. Of course, all of that is incredibly theory-driven. Those who don't accept my theoretical commitments ought not be moved. Nevertheless, I and those who share my theoretical commitments have reason to look elsewhere for a solution to these underdetermination worries.

Halpern & Hitchcock provide another potential response. On this line, what's gone wrong with the structural equation modeling consisting of the sole equation $W := S \vee B$ is that, if that model is enriched with *more* variables in a certain way, then the (HP) account's causal judgments are reversed (a cause becomes a non-cause). This might tell us that the original model was incorrect. Here are Halpern & Hitchcock (2010):

> Suppose that we have an infinite sequence of models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ such that the variables in $\mathcal{M}_i$ are $C_0, \ldots, C_{i+1}, E$, and $\mathcal{M}_{i+1, C_{i+1}=1} = \mathcal{M}_i$ (so that $\mathcal{M}_{i+1}$ can be viewed as an extension of $\mathcal{M}_i$). Is it possible that whether $C_0 = 1$ is a cause of $E = 1$ can alternate as we go through this sequence? This would indicate a certain 'instability' in the causality. In this circumstance, a lawyer should certainly be able to argue against using, say, $\mathcal{M}_7$ as a model to show that $C_0 = 1$ is a cause of $Y = 1$. On the other hand, if the sequence stabilizes, that is, if there is some k such that for all $i \geq k$; $\mathcal{M}_i$ delivers the same verdict on some causal claim of interest, that would provide a strong reason to accept $\mathcal{M}_k$ as sufficient.[19]

abstract form of a counterexample to be specified (in fact, we should be worried if our account *didn't* allow this).

[19] Halpern & Hitchcock 2010, §4.1, with notational changes

This suggests that we might want to impose a kind of *stability* criterion on a model's judgments about causation. If there is some way of enriching the variables appearing in the model which changes our account's judgments about which variable values are causally related, then the original model is not correct. Note two things about this strategy. Firstly, we could define stability only with respect to *causes*, only with respect to *non-causes*, or with respect to both. On the first approach, if our account looks at $\mathcal{M}$ and rules that $C = c$ caused $\mathbf{E} = \mathbf{e}$, then, if there exists a model $\mathcal{M}'$ with more variables than $\mathcal{M}$ such that our account looks at $\mathcal{M}'$ and rules that $C = c$ *didn't* cause $\mathbf{E} = \mathbf{e}$, then the original model was incorrect. However, if our account looks at $\mathcal{M}$ and rules that $C = c$ *didn't* cause $\mathbf{E} = \mathbf{e}$, and there is some model $\mathcal{M}'$ with more variables that $\mathcal{M}$ such that our account looks at $\mathcal{M}'$ and rules that $C = c$ *did* cause $\mathbf{E} = \mathbf{e}$, this does not mean that the original model was incorrect. That is: on the first approach, if enriching the variable set leads a cause to become a non-cause, then the original model is incorrect; whereas, the fact that enriching the variable set leads a non-cause to become a cause does not tell us that the original model was incorrect. On the second approach, if enriching the variable set leads a non-cause to become a cause, then the original model was not correct; however, just because enriching the variable set leads a cause to become a non-cause, this does not mean that the original model was incorrect. And, on the final approach, if enriching the variables leads a cause to become a non-cause or a non-cause to become a cause, then the original model was incorrect.

Secondly, the notion of stability in play here will depend upon the particulars of our account of singular causation. Therefore, if we take stability to be a necessary condition on the correctness of a model, then whether a model is correct will depend upon our account of singular causation. This is at least a *prima facie* mark against such a strategy. If we have hopes of offering a reductive analysis of causation in terms of relations of structural determination, then we ought to have hopes that we'll be able to tell a story like this: there are networks of structural determination out in the world. We can specify what it takes for the world to have a certain determination structure *without reference* to facts about which events caused which other events. Singular causation can then defined in terms of these networks of structural determination. With the stability criterion, however, we are no longer able to tell this story, and it's unclear what other reductive story we have to tell in its stead.

Halpern (ms) shows that, while including further endogenous variables in a model can lead (HP) to revise its judgment about *causes*, it cannot lead (HP) to revise its judgments about *non-causes*. He begins by defining the notion of a *conservative extension* of a causal model. $\mathcal{M}' =< \mathcal{U}', \mathcal{V}', \mathcal{E}' >$ is a *conservative extension* of $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ just in case $\mathcal{U} = \mathcal{U}'$, $\mathcal{V} \subsetneq \mathcal{V}'$, and for every $V \in \mathcal{V}$, every $\overline{\mathbf{V}}^* \subseteq \overline{\mathbf{V}} \overset{\text{def}}{=} \mathcal{V} - \{V\}$, every assignment $\overline{\mathbf{v}}^*$ to $\overline{\mathbf{V}}^*$, and every possible setting of the exogenous variables $\mathbf{u}$,

$$\mathcal{M}, \mathbf{u} \models \overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \mathrel{\Box\!\!\rightarrow} V = v \qquad \Longleftrightarrow \qquad \mathcal{M}', \mathbf{u} \models \overline{\mathbf{V}}^* = \overline{\mathbf{v}}^* \mathrel{\Box\!\!\rightarrow} V = v$$

That is: $\mathcal{M}'$ is a conservative extension of $\mathcal{M}$ just in case $\mathcal{M}'$ has all the same exogenous variables as $\mathcal{M}$, strictly more endogenous variables than $\mathcal{M}$, and $\mathcal{M}'$ entails all the same counterfactuals about the endogenous variables in $\mathcal{M}$ as $\mathcal{M}$ does. He then proves that, if the (HP) account of singular causation rules $C = c$ a non-cause of $\mathbf{E} = \mathbf{e}$ in $\mathcal{M}$, then it will rule $C = c$ a non-cause of $\mathbf{E}$ in any conservative extension of $\mathcal{M}$. That is, non-causation is, in a certain sense, stable.

This is an interesting result; however, it does not appear to be relevant to our present worries. If we thought that a causal model counts as *correct* iff all of its associated counterfactuals are true and the model is stable, in the sense that no conservative extension overturns its causal judgments, then we would think that this result means that, if we have a structural equations model all of whose counterfactuals are true, then we can trust the (HP) account's judgments of non-causation, even though we might want to remain skeptical about its judgments of causation. However, it seems clear that this is *not* the notion of stability that we're after. There may be no conservative extension, in Halpern (ms)'s sense, of a model within which our account of singular causation tells us that $C = c$ causes $\mathbf{E} = \mathbf{e}$; however, we shouldn't conclude from this that $C = c$ doesn't cause $\mathbf{E} = \mathbf{e}$.

To see this, note that the canonical representations of the neuron diagrams shown in figures 17 and 27, $\mathcal{E}_{18}$ and $\mathcal{E}_{28}$, are *not* conservative extensions of the impoverished models $\mathcal{E}_{26}$ and $\mathcal{E}_{30}$. That's because both $\mathcal{E}_{18}$ and $\mathcal{E}_{28}$ contain more *exogenous* variables than $\mathcal{E}_{26}$ and $\mathcal{E}_{30}$. Yet Halpern (ms)'s definition of *conservative extension* required that the exogenous variables are held fixed when we extend the model. So, it seems to me, the notion of *stability* that we're after is not the notion of holding true within every conservative extension. We may, of course, generalize Halpern's notion by defining a new notion, call it *conservative extension*$^+$ so that $\mathcal{M}'$ is a conservative extension$^+$ of $\mathcal{M}$ iff $\mathcal{U} \cup \mathcal{V} \neq \mathcal{U}' \cup \mathcal{V}'$, $\mathcal{U} \subseteq \mathcal{U}'$, $\mathcal{V} \subseteq \mathcal{V}'$, and all the counterfactuals true in $\mathcal{M}$ are true in $\mathcal{M}'$. However, then Halpern (ms)'s result no longer holds—(HP) may rule $C = c$ a non-cause of $\mathbf{E} = \mathbf{e}$ in $\mathcal{M}$ yet rule $C = c$ a cause of $\mathbf{E} = \mathbf{e}$ in a conservative extension^{+} of $\mathcal{M}$ (see footnote 16). So I don't believe that this result is particularly relevant to the question of how to avoid underdetermination worries.

In the absence of a result like Halpern (ms)'s, if we accept stability as a criterion of correctness, then figuring out whether $C = c$ caused $\mathbf{E} = \mathbf{e}$ becomes incredibly complex. It is not enough to recognize some network of structural determination relations out in the world and apply some simple counterfactual test. One must additionally explore all networks of structural determination which include some additional exogenous variables. But there will always be myriad additional, not-too-unnatural variables which may be included in a model. For any exogenous variables in a given model, we can simply consider the variables which structurally determine them; and then consider the variables which structurally determine the variables which structurally determine the exogenous variables, and so on and so forth. On an account like this, without a result like Halpern (ms)'s, it becomes mysterious how we are often able to attain causal knowledge so effortlessly.

Halpern & Hitchcock (2010) suggest another kind of stability that a model might have: not stability with respect to our account of singular causation's judgments; but rather stability with respect to certain features of the structural equations model itself. Their suggestion is that

> ...adding additional variables to a model will not affect the relations of [singular] causation that hold in the model unless the addition of those variables changes the 'topology' of the model.

The suggestion is, then, that in order for a model to be correct, it must entail only true counterfactuals, and, moreover, there must be no other model including all the variables from the first, which entails only true counterfactuals, and which has a relevantly different topological structure than the original model.

Here is a way of precisifying this idea. Say that a parent-child relationship between the variables C and P in a causal model—the relationship represented in a causal graph by an arrow with its tail at P and its tip at C—is *autonomous* if there is no other parent of C, P', and no omitted variable O, such that, by including the omitted variable O in our model, we get a correct causal model within which P and P' are parents of O, and O is an ancestor of C.

> *Autonomy*: $P \rightarrow C$ is an *autonomous* structural determination relation in a causal model $\mathcal{M}$ iff there is no correct causal model $\mathcal{M}'$ such that

$\mathcal{V} \subsetneq \mathcal{V}'$ and, in $\mathcal{M}$, $P, P' \in \mathbf{PA}(C)$, and in $\mathcal{M}'$, $P, P' \in \mathbf{PA}(O)$ and $O \in \mathbf{AN}(C)$.[20]

That is: if there's a way of tying together multiple of the arrows leading from $\mathbf{PA}(C)$ to C, so that some of C's parents only influence C via some intermediary variable O, then these arrows do not represent *autonomous* structural determination relations. Naturally, we must rule out gerrymandered omitted variables to keep this definition from ruling every structural determination relation non-autonomous. But some such restriction on the kinds of variables that get to enter into a structural equations model is going to be needed if there's any hope of using these models to give an account of causation.

If we then say that a structural equations model is only correct if all of its structural determination relations are *autonomous*, then we can say that the system $\mathcal{E}_{30}$ correctly represents the neuron diagram displayed in figure 27, whereas $\mathcal{E}_{26}$ does *not* correctly represent the neuron diagram displayed in figure 17—for the structural determination relation $c \to g$ is not autonomous from the structural determination relation $e \to g$. We can tie these two structural determination relations together by including the variable f in our model. With f included we get the system of equations $\mathcal{E}_{25}$, in which $c, e \in \mathbf{PA}(f)$ and $f \in \mathbf{PA}(g)$. Adopting *autonomy* as a necessary condition for the correctness of a causal model means rejecting **Functional Composition Preserves Correctness**, since functional composition can remove variables lying along the joint path by which two ancestors influence their descendant. We can, however, retain **Removing Exogenous Variables Perserves Correctness**.

For my money, this final suggestion is the most promising—*i.e.*, singular causation can only be uncovered in structural equations models all of whose structural determination relations are *autonomous*. Of course, this suggestion could very well be combined with the suggestion that two variable values are causally related iff there is *some* structural equations model within which the account rules them to be causally related. Indeed, this dual approach is the one that I adopt in chapter 5. However, I will not defend it any further here. The proof of this pudding lies in the eating of chapter 5. Here, I am content to have 1) argued that if we think that the truth conditions for structural equations models are given by causal counterfactual conditionals—or if we simply accept the three principles laid down in §4.4—then relations of singular causation are underdetermined by a correct structural equations model, and 2) outlined

[20] '$\mathbf{AN}(C)$' is a vector of C's ancestors.

some general strategies which we may adopt to deal with this problem, and discussed some of the advantages and pitfalls of these strategies.

CHAPTER 5

A THEORY OF SINGULAR CAUSATION

5.1 Introduction

In chapter 3, I argued that an understanding of structural equations models according to which their correctness conditions are provided by causal counterfactual conditionals is incapable of securing the independent manipulability of the structural determination relations in a correct structural equations model—a property known as *modularity*. In contrast, I presented an alternate understanding of the content of a structural equations model, dubbed the 'nomic sufficiency' understanding, according to which a structural equations model is correct only if, for every endogenous variable in the model, there exists a set of possibilities meeting certain constraints, and throughout which that endogenous variable's structural equation is descriptively adequate. It was demonstrated that this account entails that structural determination relations are independently manipulable, that it provides a way of understanding the technical notion of an *intervention*, as well as a criterion for distinguishing ways of setting the values of variables which do from those which do not constitute interventions, and that it allows correct structural equations models to provide a reductive semantics for causal counterfactual conditionals.

In chapter 4, I introduce the additional requirement that the structural determination relations in a correct structural equations model be *autonomous*. What it is for a structural determination relation between a parent variable $P \in \mathbf{PA}(C)$ and its child C to be autonomous is for there to be no way of including omitted variables in the model which delivers a correct structural equations model according to which two parents $P, P' \in \mathbf{PA}(C)$ in the original model are both parents of an omitted variable, O, which is an ancestor of C in the new model. If the structural determination relations in a structural equations model are autonomous, then each arrow in the causal graph represents a structural determination relation which operates independently of all of the others.

Here, I will use the relations of structural determination codified in a correct structural equations model to provide an account of singular, token, or actual causation.[1] Key to my account will be the distinction between events in which a variable value *changes* and those in which a variable value remains constant. However, given the way that I defined variables in chapter 3—as a mapping from possible worlds to the real line—variable values are not capable of changing over time within a world. I will therefore introduce *dynamic* variables—mappings from world-time pairs to the real line.

I will then introduce *dynamic* structural equations models, which are capable of modeling structural determination relations between dynamic variables, determination relations which endure through time. In rough outline, the view to be advanced here is that such structural determination relations provide the pathways along which causal influence propagates, and that causation is a form of counterfactual dependence in a correct structural equations model—sometimes, and in a certain way, transitively closed. In §5.2, I will introduce dynamic structural equations models and explain how they allow us to draw the important distinction between actual and merely hypothetical interventions. In §5.3, I will develop my account of singular causation. On this account, there are important causal differences between events in which variable values *change* and those in which they remain constant. For this reason, I will dub this account the *dynamic* counterfactual account of singular causation. Throughout §5.3, I will demonstrate how the dynamic account handles cases of early and late preemp-

[1] I take these terms to be synonymous. For the sake of consistency, I'll stick to 'singular' throughout.

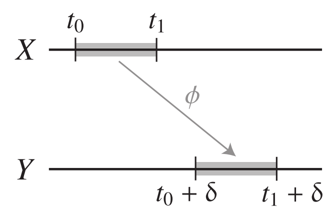


Figure 31: Structural Determination between dynamic variables. The value of a dynamic variable at one time can determine the value of another dynamic variable at a distinct time. In the figure above, Y is determined by X during the interval $[t_0 + \delta, t_1 + \delta]$, and X determines Y during the interval $[t_0, t_1]$. This structural determination relation is represented by the structural equation $Y_t = \phi(X_{t-\delta}), \ \ \forall t \in [t_0 + \delta, t_1 + \delta]$.

tion (§5.3.1.1), preemptive prevention (§5.3.2.1), transitivity violations (§5.3.5.2)—including cases of switching (§5.3.8)—short circuits (§5.3.7), symmetric overdetermination (§5.3.9), preemptive double prevention (§5.3.10), and trumping (§5.3.11).

5.2 Dynamic Structural Equations Models

As I said in chapter 3, the variables which enter into structural determination relations with one another are standardly understood to be partial functions from the set of possible worlds to the real line.[2] This has the consequence that variable values cannot change over time at a world. It will be important to my account of singular causation that the variables entering into relations of structural determination with one another *can* change their values over time. So let us start thinking about variables, not as partial functions from possibile worlds $\omega \in \Omega$ to the real line, but rather as partial functions from world-time *pairs*, $< \omega, t > \in \Omega \times \mathbb{T}$ (where $\mathbb{T}$ is the set of times t), to the real line, $\mathbb{R}$. We can call variables like these *dynamic* variables. Call variables from worlds to $\mathbb{R}$ *static* variables. Just as the value of static variables can vary over different worlds, the values of dynamic variables can vary over time within a world. For any dynamic variable V, I'll use '$V_{\omega,\,t}$' to denote the value of V at world-time pair $< \omega, t >$, just as 'V_ω' denoted the value of V at world ω. As before, I will use '$V_t = v$' to denote the proposition that V takes on the value v at t (*i.e.*, the set of worlds ω such that $V_{\omega,t} = v$).

For illustration, consider a dynamic system of structural equations like the following:

[2] At least, this is how they are defined *when* they are explicitly and rigorously defined, which is not as often as one which have hoped—see Hall (2007).

$$\mathcal{E}_{32} = \left(\begin{array}{ll} Y_t := \phi_Y(X_{t-\delta}) & \forall t \in [t_0, t_1] \\ Z_t := \phi_Z(X_{t-\delta}, Y_{t-\delta}) & \forall t \in [t_0, t_1] \end{array} \right)$$

$$\begin{tikzcd} X \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\ & Z & \end{tikzcd}$$

Figure 32

Here, the value of Y at t is structurally determined by the value of X at $t-\delta$, for any t in the interval $[t_0, t_1]$. And the value of Z at t is structurally determined by the value of both X and Y at $t-\delta$, for any t in the interval $[t_0, t_1]$. (Notice that, here, variables at distinct time indices can enter into structural determination relations with one another. For instance, even though Y is structurally determined by X, it takes δ seconds for a change in the value of X to affect the value of Y. Y's value at any time is structurally determined not by X's value at that time, but rather by X's value δ seconds prior. Similarly, Z's value at any time is structurally determined by the values X and Y had δ seconds prior. See figure 31.) If we take each of the variables appearing in these equations to be static variables, then, for every $t \in [t_0, t_1]$, we will have a structural equation of the form $Y_t := \phi_Y(X_{t-\delta})$. Since, on this understanding, for every $t, t' \in [t_0, t_1] (t \neq t')$, Y_t, $Y_{t'}$, $X_{t-\delta}$, and $X_{t'-\delta}$ are distinct variables, this means that the structural determination relation between Y_t and $X_{t-\delta}$ must be distinct from the structural determination relation between $Y_{t'}$ and $X_{t'-\delta}$. Since there are continuum many times in $[t_0, t_1]$, there will be continuum many static variables in our model, and therefore continuum many structural determination relations. On the other hand, if we treat X and Y as dynamic variables, then we need have only a single structural determination relation between X and Y which persists from t_0 to t_1.

What we mean by 'structural determination relation' thus depends upon whether we take the variables to be static or dynamic. If we confine our attention to static variables, then the term 'structural determination relation' will not pick out something which persists through time. If we confine our attention to dynamic variables, then 'structural determination relation' will pick out something which persists through time. Nevertheless, so long as the values of the static variables depend solely upon the state of the world at a certain specific time—as my discussion in chapter 3 presupposed that they did—we can straightforwardly translate talk about structural determination relations between dynamic variables into talk about structural determination relations between (perhaps continuum many) static variables. What it is for a dynamic structural determination relation between the dynamic variables Y and X to persist from t_0 to t_1 is just for there to be continuum-many static structural determination relations between the static Y_t and $X_{t-\delta}$, for all $t \in [t_0, t_1]$.[3] So we can translate everything that I have to say about dynamic structural determination between dynamic variables into talk about static structural determination between static variables. So translated, the account I offered in chapter 3, together with the *autonomy* condition from chapter 4, will provide the correctness conditions for dynamic structural equations models.

5.2.1 Actual versus Hypothetical Interventions

In chapter 3, I noted that the nomic sufficiency account of structural equations models allowed us to give an account of which ways of setting a variable value count as *interventions*. One advantage of the move to dynamic causal models which is worth mentioning is that it allows us to easily distinguish between two kinds of interventions: *actual* and *hypothetical* interventions. That is, the interventions that are actually performed and those which are not actually, but could possibly have been, performed. In the case of hypothetical interventions, we consider a non-actual world in which the structural equation ϕ_V is not in force, despite all the other structural equations in the model remaining in force, as in figure 33. In the case of actual interventions, the actual

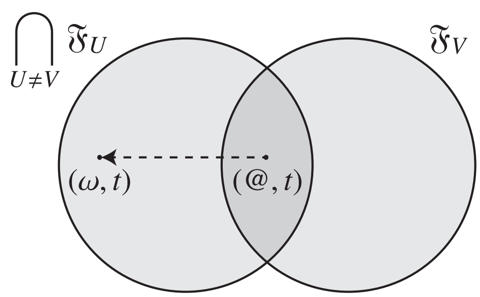


Figure 33: A hypothetical intervention

world's history takes it on a path from a time during which all the equations in $\mathcal{E}$ are in force to a time during which ϕ_V is not in force, but all the equations in $\mathcal{E} - \{\phi_V\}$ are, as shown in figure 34.

[3] If the value of Y at t_1 is structurally determined by *all* values had by X throughout $[t_0, t_1]$, as in $\mathcal{E}_{36}$ below, then there is a single structural determination relation between the static Y_{t_1} and the continuum-many static variables X_t, for all $t \in [t_0, t_1]$.

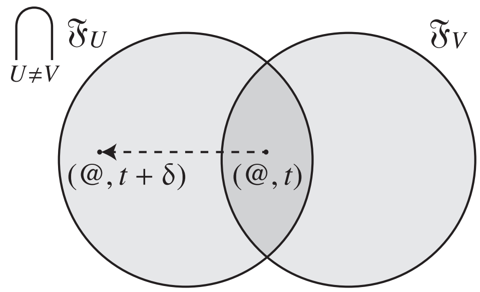


Figure 34: An actual intervention

While the distinction between actual and merely hypothetical interventions is quite natural, it is difficult to make sense of actual interventions if we think of structural equations models as relating static variables. That's because, if we are dealing with variables V, U which only take on a single value at any possible world, then either V is actually structurally determined by U, in which case there is no actual intervention, or V is not actually structurally determined by U, in which case it's unclear why we should think that any event at the actual world which determines the value of V should count as destroying or removing a structural determination relation between V and U—*ex hypothesi*, there is not and never was any such dependence relation. I think the best thing to say about actual interventions in a framework that deals exclusively with static variables is that an intervention destroyed U's structural determination of V just in case, had the intervention not taken place, U would have structurally determined V. However, this counterfactual definition of an actual intervention runs into the same kind of preemption worries that the counterfactual analysis of causation runs into. Simply because I is an actual intervention on the value of a variable V which destroys its determination by U, this doesn't mean that, had I not occurred, V would have been determined by U, since it could be that, had I not occurred, *another* intervention I' would have occurred, which would have destroyed the determination of V by U. I could continue adding epicycles, but there's no reason to. My point isn't that there's *no* way to make sense of actual interventions without dynamic structural equations models; it is rather merely that it's difficult to make sense of actual interventions using static structural equations models, and that it isn't nearly as difficult to make sense of them using dynamic models.

5.3 A Dynamic Theory of Singular Causation

If we think of a causal model as relating these *dynamic* variables, then it describes a determination structure which persists through time. On this way of understanding structural determination, we can begin to take much more seriously the understanding of structural determination which says that E is determined by $\mathbf{PA}(E)$ just in case wiggling the variables in $\mathbf{PA}(E)$ wiggles E, since, on this understanding of variables, $\mathbf{PA}(E)$ and E are actually capable of *wiggling*. The values of $\mathbf{PA}(E)$ can actually *change*, and if E is structurally determined by the variables in $\mathbf{PA}(E)$, then this change in the values of $\mathbf{PA}(E)$ can bring about a change in the value of E. In the previous sentence, I said 'bring about', but I might just as well have said 'cause'. Let me go ahead and say it, then: one thing, at least, that can cause and be caused is changes in the values of dynamic variables. The structural determination relations represented in a correct dynamic structural equations model provide the pathways along which this causal influence propagates.

On my view, an understanding of all the complexities of singular causation, and our thought and talk about it, requires distinguishing events in which a variable value *changes* from those in which its value stays constant. Events of the first kind are exciting; they are the events that usually attract our attention. Events of the second kind are boring; they rarely catch our attention, but they are no less fit to enter into causal relations for all that. All causation reduces to a species of causal counterfactual dependence, but causal relations between exciting events require different treatment than those relating boring events. While causal relations between exciting events are always transitively closed (or near enough—more on that below), those relating boring events are not. Because this account places so much weight on the kinds of changes that variable values undergo or fail to undergo, and the times at which they (fail to) undergo them, I will dub it the *dynamic* counterfactual theory of singular causation.

To make explicit what was implicit in the foregoing: On the dynamic theory, the causal relata are the events of variables taking on or changing their values at certain times. This cuts the causal relata somewhat thin—the event of John's saying 'hello' will be distinct from the event of John's saying 'hello' loudly, since a binary variable which tells us whether John says 'hello' will be distinct from the variable which tells us how loudly John says 'hello'—which will, of course, be distinct from the variable which tells us both *whether* John says hello, and, if he said 'hello', how loudly he said it. (For this reason, it allows for the existence of high-level events, in the sense of chapter

2. Moreover, simply because there is a structural determination relation between the low-level variable describing the fundamental physical state of Gliese 163c's orbit and the low-level variable describing the fundamental physical state of the U.S. economy, this does not mean that there will be a structural determination relation between a variable describing whether Gliese 163c has orbited and a variable describing whether the U.S. economy has stagflated (see figure 2). For this reason, so long as we are allowed to feed high-level variables into the dynamic account, it will entail the thesis I called 'causal emergentism' in chapter 2.)

5.3.1 Productive Causation

Some notation: I'll represent a change in a variable V's value at time t with 'ΔV_t'. I'll represent a change in a variable V's value, from v_0 to v_1, at time t with '$\Delta(v_0, v_1)_t^V$'. And I'll use '$\mathcal{C}$' to represent the singular causal relation. If there is a network of structural determination relations described by a correct structural equations model $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$, with $C, E \in \mathcal{U} \cup \mathcal{V}$, then we can lay down the following sufficient condition for a change in C's value, from c_0 to c_1, at time t causing a change in E's value, from e_0 to e_1, at time t', at world ω, according to the model $\mathcal{M}$:

(Prod 1) $$\mathcal{M}, \mathcal{U}_{\omega,[f_E(t'),t']} \models \neg\Delta C_t \mathrel{\Box\!\!\rightarrow} \neg\Delta E_{t'}$$

That is, if $\mathcal{M}$, together with a specification of the values of the variables in the model throughout the relevant time period at ω, entails that, had C not changed its value at t, then E wouldn't have changed its value at t', then according to $\mathcal{M}$, C's change in value caused E's change in value, $\mathcal{C}(\Delta C_t, \Delta E_{t'})$. The *relevant* time period is the time period during which E's ancestors could have had an effect upon E. Each structural determination relation between a variable and its children will involve some time lag (in the case of instantaneous causation, this time lag will be zero). Add up the time lags along every directed path between E and each of its exogenous ancestors. Let $l(E)$ be the longest such time lag. Then, let $f_E(t)$ be $t - l(E)$. Then, the relevant time period is the period $[f_E(t'), t']$.

We should evaluate this counterfactual as follows: in order to satisfy the antecedent, we mutilate the model, severing the determination of C by $\mathbf{PA}(C)$, and make it the case that C's value doesn't change at t. That means that C should stay at its original value c_0 at t. We then see whether, according to the structural equations in $\mathcal{E} - \{\phi_C\}$, given the actual values of the variables during the period $[f_E(t'), t']$, E's value changes to e_1 at t'. There are three ways this could fail to happen: either 1) E could fail to change its value to e_1 *at* t', changing its value either earlier or later instead, or 2) E could fail to *change its value* to e_1 at t', remaining at e_0 instead, or 3) E could fail to change its value *to* e_1 at t', changing its value to something else instead. In any of these cases, ΔC_t is a cause of $\Delta E_{t'}$.

It may seem that to say this is not yet to say very much—it is, after all, merely to say that a certain kind of counterfactual dependence in a correct structural equations model is sufficient for singular causation. Most everyone in the counterfactual tradition is willing to accept that causal counterfactual dependence is sufficient for causation. So, if a correct structural equations model succeeds at its job of providing a semantics for causal counterfactuals, then (Prod 1) should be true. Note, though, that this counterfactual appeals only to the changes in variables values at times, and not to the occurrence or non-occurrence of events. In order to evaluate this counterfactual, we don't have to say anything about whether, if C hadn't changed its value at t, there would have been a different event in its place; nor do we have to say anything about which possibility in which C's value doesn't change is the possibility most similar to actuality. Though we've said that the causal relata are the events of variable values taking on such-and-such values, or changing their values at such-and-such time, in order to say this, we don't have to say anything about the modal profile of these events. We just evaluate the counterfactual in the straightforward way outlined in the previous chapter: sever the connections between C and it's causal parents at t, set its value to c_0, and solve for the values of all the other variables in the model according to their structural equations, given the values of their parent variables during the period $[f_E(t'), t']$.

We can go further by considering another kind of counterfactual dependence—what we can call *local* counterfactual dependence. Local counterfactual dependence is not counterfactual dependence of $\Delta E_{t'}$ on ΔC_t within a *full* network of structural determination. Rather, it is counterfactual dependence of $\Delta E_{t'}$ on ΔC_t within the local determination structure relating E and its parent variables. $\Delta E_{t'}$ locally depends upon ΔC_t at ω according to $\mathcal{M} =< \mathcal{U}, \mathcal{V}, \mathcal{E} >$ (with $\phi_E \in \mathcal{E}$) just in case (Prod 2).

(Prod 2) $$\phi_E, \mathbf{PA}(E)_{\omega,[f_E(t'),t']} \models \neg\Delta C_t \;\Box\!\!\rightarrow \neg\Delta E_{t'}$$

That is, if the isolated structural equation ϕ_E, together with the values of E's parent variables throughout the period $[f_E(t'), t']$ entails that, had C not changed its value at t, then E wouldn't have changed its value at t', then $\Delta E_{t'}$ *locally* counterfactually depends upon ΔC_t at ω according to $\mathcal{M}$. (Note that ϕ_E could only entail this if $C \in \mathbf{PA}(E)$, so a change in variable value can only locally counterfactually depend upon the values of its parent variables in some correct structural equations model.) This counterfactual is to be evaluated in the same way as the previous one, except with our attention confined to E's structural equation alone—it is true so long as, were C to remain at its previous value at t, and were all of the variables in $\mathbf{PA}(E)$ to be held fixed at their actual values during $[f_E(t'), t']$,[4] ϕ_E tells us that E would have either 1) never changed its value, or 2) changed its value earlier or later, or 3) changed its value to something else.

Below, I will explain what looking at local counterfactual dependence gains us. For now, let me just state that, if a change in E's value at t' counterfactually depends, either globally or locally, upon a change in C's value at t in some correct dynamic structural equations model, then ΔC_t is what I will call a *productive* cause of $\Delta E_{t'}$.

(Prod) $$\exists \mathcal{M} \ : \ (\text{Prod 1}) \text{ or } (\text{Prod 2}) \quad \Longrightarrow \quad \mathcal{C}(\Delta C_t, \Delta E_{t'})$$

The reason for the 'productive' modifier is that I do not think that changes in variable values are the only kinds of events that can be causally related. A variable's static value at a time can also cause and be caused. I'll have more to say about those kinds of causal relata in the following sections. Right now, I just want to focus on productive causation.

5.3.1.1 Preemption

Billy and Suzy stand in front of a window with rocks in hand. Billy is a bit closer than Suzy. Both Suzy and Billy throw rocks at the same speed, and both have excellent aim. If either of them throw their rocks at the window, then the window will shatter. In this case, the structural determination relations described by the system of equations $\mathcal{E}_{35}$, shown in figure 35, will be in force. (In figure $\mathcal{E}_{35}$, $\delta > \epsilon$). Suppose that Billy never throws his rock, but that, at some time $t - \delta$, Suzy throws her rock at the window. This means that, at $t - \delta$, S changes its value from 0 to 1. So, at t, the window will break, so at t, W will change its value from 0 to 1. According to the account given here, $\Delta(0,1)^S_{t-\delta}$ is a productive cause of $\Delta(0,1)^W_t$ (either disjunct of (Prod) will get this result).

[4] Holding fixed does not mean holding *static*. If the variables in $\mathbf{PA}(E)$ change their values during $[f_E(t'), t']$, then such changes should be held fixed when evaluating the counterfactual.

$$\mathcal{E}_{35} = (W_t := S_{t-\delta} \vee B_{t-\epsilon} \quad \forall t \in \mathbf{T})$$

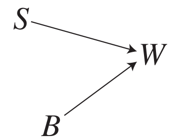


Figure 35

$$\mathcal{E}_{36} = \begin{pmatrix} W_t & := S_{t-\delta} \vee B_{t-\epsilon} & \forall t \in [t_0, t'] \\ B_{t'-\epsilon} & := \forall t \in [t_0, t'-\epsilon] \ \overline{S}_t & \end{pmatrix}$$

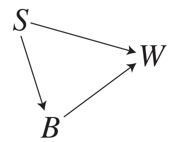


Figure 36

Suppose, on the other hand, that both Suzy and Billy are determined to have the window shatter before t'. Billy will let Suzy shatter the window, but if she waits too long, then he will have to throw in order to have the window break in time. This is a case of *early preemption*. In this case, $\mathcal{E}_{36}$, shown in figure 36, will be in force. (In $\mathcal{E}_{36}$, $\delta > \epsilon$, and, in the bottom equation, the quantifier is a part of B's structural equation—the state of B at $t' - \epsilon$ is determined by the state of S at all times between t_0 and $t' - \epsilon$.) If Suzy throws before $t' - \epsilon$, then Billy will not throw. However, if she doesn't throw before $t' - \epsilon$, then Billy will throw. Suppose that Suzy throws at $t' - \delta$, and so the window shatters at t'. If we evaluate the counterfactual

$$\neg\Delta S_{t'-\delta} \mathrel{\Box\!\!\rightarrow} \neg\Delta W_{t'}$$

with respect to the system of equations $\mathcal{E}_{36}$ and the values of S, B, and W during the relevant time period, it will tell us that the counterfactual is false. $\Delta W_{t'}$ does not counterfactually depend upon $\Delta S_{t'-\delta}$. Had Suzy not thrown at $t' - \delta$, then Billy would have thrown at $t' - \epsilon$, and the window would have shattered at t' just the same. However, $\Delta W_{t'}$ does *locally* counterfactually depend upon $\Delta S_{t'-\delta}$. If we look just at W's structural equation, given the actual values of S and B from t_0 to t', $\Delta S_{t'-\delta} \mathrel{\Box\!\!\rightarrow} \neg\Delta W_{t'}$ is true. Given that Billy didn't actually throw his rock, had Suzy not thrown hers, the window wouldn't have shattered. So (Prod 2) tells us that Suzy's throw was a productive cause of the window's shattering.

Suppose, finally, that the system of structural equations $\mathcal{E}_{35}$ is in force, and both Suzy and Billy throw their rocks at the same time—*i.e.*, S and B change their value from 0 to 1 at t. Then, $\mathcal{E}_{35}$ tells us that, had Suzy not thrown her rock at t, the bottle would not have shattered at $t + \delta$. Rather, it would have shattered at $t + \epsilon$. So the counterfactual

$$\neg \Delta S_t \mathrel{\Box\!\!\rightarrow} \neg \Delta W_{t+\delta}$$

is true. So (Prod) tells us that Suzy's throwing the rock caused the bottle to shatter.

That's a case (commonly called a case of *late preemption*) that Mackie's 1965 account fails to get right, since Billy's throw is an INUS condition for the bottle shattering. It's a case that Lewis's 1973 counterfactual account was never able to get quite right, since the bottle would have shattered even in the absense of Suzy's throw and there are no events intermediate between the throw and the shattering such that the shattering counterfactually depends upon them. It's one of the main problem cases which motivated the epicycles of quasi-dependence[5] and Lewis's 2000 *influence* account of causation.

Here, we were able to get the result that Suzy's throw caused the window to shatter rather effortlessly. We did so by playing on a feature of the case which at least one undergraduate seems to pick up on whenever the case is taught: while the window would have still shattered if Suzy hadn't thrown, it would have shattered a little bit later. What we tell that undergraduate is what Lewis told us: that they are playing a dangerous game, making the events this fragile. What's good for the goose is good for the gander. If the window's shattering could have so easily not occurred, then Suzy's throw could also have easily not occurred; she could have thrown her rock from a slightly different angle, or with one eye closed, or with a smirk. And if it is so easy to destroy the event of Suzy's throwing the rock, then surely one of *these* is the closest possibility in which the throw doesn't occur. But the window would have still shattered at the same time if Suzy had thrown her rock with a smirk. And, even if not, by making the window's shattering this fragile, we've opened the door to a host of non-causes. Billy and his rock exert a gravitational pull on the window, so, if the shattering of the window becomes too fragile, it may turn out that Billy caused the window to shatter after all! And even if it's hard not to feel happy for poor Billy, finally getting to shatter that window, we shouldn't let our emotions cloud our judgment, the hard truth is that Billy's throw did nothing to bring about the shattering of the window.

In contrast, all the above account talks about is the values of variables at times. It says that, were S to have stayed at 0 at t, W wouldn't have flipped to 1 at $t + \delta$. But saying this doesn't commit us to anything about the modal profile of the *event* of Suzy's throw or the window's shattering. The window could have shattered ear-

[5] See Lewis (1986a, postscript E)

lier or later than it actually did; it could have been caused by Billy's throw; it could have been shattered by a baseball rather than a rock. Even though (Prod) appeals to counterfactuals, the counterfactuals aren't of the form '$\neg O(C) \;\Box\!\!\rightarrow\; \neg O(E)$', so we don't have to futz with questions about what it takes for an event to fail to occur;[6] similarly, since we're not using the standard Lewis/Stalnaker semantics for evaluating these counterfactuals, but rather utilizing the methods afforded us by structural equations models, we don't have to futz with questions about which of the ways that Suzy's throwing the rock could fail to occur is most similar to the actual world. (Does Suzy throw something else? Does she throw it *at* something else? Does she not throw anything at all?)[7]

Paul & Hall (2013) say the following about this kind of strategy:

> ...it seems easy enough to construct late preemption examples in which, had the cause not occurred...the effect would have occurred at exactly the same time.[8]

I concur. However, in these cases, a different system of equations will be needed to correctly describe the structural determination relations amongst the variables. For instance, suppose that both Suzy and Billy stand equidistant from the window, and both throw their rocks at the same time, but that Suzy's rock, if thrown, exerts a retarding force on Billy's rock which slows its velocity. This will prevent Billy's rock from reaching the window before Suzy's rock has already shattered it. However, had Suzy not thrown, her rock would not have exerted the retarding force, and Billy's rock would have shattered the window at exactly the same time. In this case, the structural equation $W_{t+\delta} := S_t \vee B_t$ will not be in force, since, according to this structural equation, the time it takes Billy's rock to reach the window is unaffected by whether Suzy has thrown. In this case, there will be a true structural equations model according to which whether the window shatters at a given time is determined by whether Suzy or Billy has thrown in the recent past, and also upon the velocity of Billy's rock (V); the velocity of Billy's rock, in turn, is determined both by whether Billy has thrown and by whether Suzy has thrown—see figure 37. But now, even though the window's shattering, ΔW, doesn't counterfactually depend upon Suzy's throw, ΔS, ΔW does

[6] *Cf.* Paul (2000, 2004) and Paul & Hall (2013, p. 237 ff.)

[7] Closely related questions will be relevant to the question of whether W is structurally determined by S, but they will be questions about the modal profile of the variable S, and not the modal profile of the event of Suzy's throw.

[8] Paul & Hall (2013, p. 107).

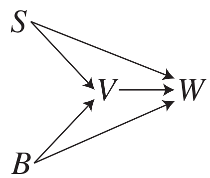


Figure 37

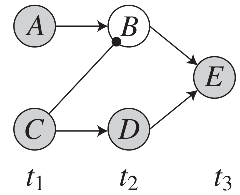


Figure 38

$$\mathcal{E}_{39} = \begin{pmatrix} B_t := A_{t-1} \wedge \overline{C}_{t-1} & \forall t \in \mathbf{T} \\ E_t := B_{t-1} \vee C_{t-2} & \end{pmatrix}$$

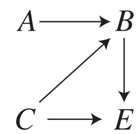


Figure 39

locally counterfactually depend upon ΔS. Holding fixed that Billy actually threw his rock and holding fixed the rock's actual velocity, ϕ_W will tell us that, had Suzy not thrown her rock, the window wouldn't have shattered when it did. So, according to (Prod), Suzy's throw productively causes the window's shattering in this scenario as well.

Not all cases of preemption are cases in which the preempting cause is a parent of its effect. For instance, consider the neuron diagram displayed in figure 38. There, C's firing caused E's firing, even though, had C not fired, E would still have fired. However, in the first place, in the canonical structural equations model, C's firing productively caused D's firing, and D's firing productive caused E's firing; for this reason, it will end up following, on my account of the transitivity of causation, that C's firing caused E's firing (see §5.3.5). Moreover, we needn't even appeal to transitivity to get the verdict that C's firing caused E's firing. That's because the account merely requires local counterfactual dependence in *some* correct structural equations model. The structural equations model shown in figure 39 meets all the conditions given in chapter 3, and all of the structural determination relations in that model are autonomous. So it correctly models the neuron diagram in figure 38. And, in this structural equations model, ΔE_{t_3} locally counterfactually depends upon ΔC_{t_1}.

5.3.2 Preventive Causation

Changes in variable values are particularly exciting and noteworthy events, but they are not the only kinds of events when enter into causal relations. Less exciting events may cause and be caused as well. For instance, consider a structural determination network described by the following structural equation:

$$Z_t := X_{t-\delta} \vee Y_{t-\delta} \qquad \forall t \in \mathbf{T} \qquad\qquad X \longrightarrow Z \longleftarrow Y$$

where X, Y, and Z are each binary variables. This structural equation says that, for any time $t \in \mathbf{T}$, Z will be 1 iff either X or Y was 1 δ seconds prior. Now, suppose that, prior to t_0, Y is 1 and X is 0. Then, at t_0, Y changes its value to 0 and X changes its value to 1. At time $t_0 + \delta$, Z remains at 1. In that case, the change in X's value at t_0, $\Delta(0,1)^X_{t_0}$, *caused* Z's value to remain at 1 at $t_0 + \delta$, the event that I'll denote '$1^Z_{t_0+\delta}$'. But, while the cause is an exciting or interesting event—X's value changes—the effect is a boring event—Z's value does not change. I'm going to call causal relations like this, causal relations in which the cause is exciting, the effect boring, *preventive causal relations*. Here's a sufficient condition for a change in the value of a variable C at t counting as a preventive cause of the value of a variable E at t', at world ω, according to the model $\mathcal{M}$:

(Prev 1) $$\mathcal{M}, \mathcal{U}_{\omega,[f_E(t'),t']} \models \neg\Delta C_t \;\Box\!\!\rightarrow E_{t'} \neq E_{\omega,t'}$$

That is: if, according to a correct system of structural equations and the values of the variables throughout the period $[f_E(t'), t']$, were C to not change its value at t—that is, were it to stay fixed at its original value—then E would have taken on a different value at t'.

As with productive causation, *local* counterfactual dependence is also sufficient for ΔC_t to be a preventive cause of $E_{t'}$—where $E_{t'}$ locally depends upon ΔC_t at ω according to $\mathcal{M} = <\mathcal{U}, \mathcal{V}, \mathcal{E}>$ (with $\phi_E \in \mathcal{E}$) just in case (Prev 2).

(Prev 2) $$\phi_E, \mathbf{PA}(E)_{\omega,[f_E(t'),t']} \models \neg\Delta C_t \;\Box\!\!\rightarrow E_{t'} \neq E_{\omega,t'}$$

That is: holding fixed the values of E's parents other than C, if it's the case that, were C to not change its value, E would not taken on a different value at t', then ΔC_t is a preventive cause of $E_{t'}$.

Either way, ΔC_t is a preventive cause of $E_{t'}$. That is: ΔC_t is a preventive cause of $E_{t'}$ at world ω if $E_{t'}$ counterfactually depends, either globally or locally, upon ΔC_t in some structural equations model which is correct at ω.

(Prev) $$\exists \mathcal{M} : \text{ (Prev 1) or (Prev 2)} \quad \Longrightarrow \quad \mathcal{C}(\Delta C_t, E_{t'})$$

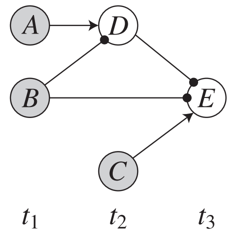


Figure 40

$$\mathcal{E}_{41} = \begin{pmatrix} D_t := A_{t-1} \wedge \overline{B}_{t-1} & \\ E_t := C_{t-1} \wedge \overline{B}_{t-2} \wedge \overline{D}_{t-1} & \forall t \in \mathbf{T} \end{pmatrix}$$

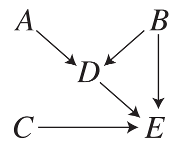


Figure 41

5.3.2.1 Preemptive Prevention

For an example of preventive causation in which we have local counterfactual dependence without global counterfactual dependence, consider the neuron diagram shown in figure 40. This neuron diagram provides a case of preemptive prevention.[9] Intuitively, B's firing, and not A's firing, prevented E from firing—even though, had B not fired, A's firing *would* have prevented E from firing. (Prev) can accommodate this intuition. The structural determination relations in this neuron diagram can be represented by $\mathcal{E}_{41}$, shown in figure 41. D will fire at t just in A fires one second prior and B doesn't fire one second prior. And E will fire just in case C fires one seconds prior, B doesn't fire two seconds prior, and D doesn't fire one second prior. The full network of structural equations tells us that, had B not fired, E still wouldn't have fired, since A would have kept it from firing. Nevertheless, looking just at E's structural equation, and holding fixed the actual values of E's parents during the relevant time period, had B not fired, E would have fired. So, E's failure to fire *locally* counterfactually depends upon B's firing. So B's firing is a preventive cause of E's failure to fire, according to (Prev). Similarly, (Prev) tells us that, had B not fired, then A *would* have prevented E from firing.

We will get the wrong verdict here if we model the neuron diagram with $\mathcal{E}_{42}$, shown in figure 42. Given this model, the dynamic account tells us, wrongly, that neither A's firing nor B's firing caused E to remain dormant.[10] However, in this

[9] See McDermott (1995), Collins (2004)

[10] Though, it would tell us that their disjunction caused E to remain dormant—see §5.3.9.

$$\mathcal{E}_{42} = \left(E_t := C_{t-2} \wedge \overline{B}_{t-2} \wedge \overline{A}_{t-2} \quad \forall t \in \mathbf{T} \right)$$

$$\begin{tikzcd} A \arrow[dr] & B \arrow[d] \\ C \arrow[r] & E \end{tikzcd}$$

Figure 42

model, the structural determination relation between A and E is not autonomous from the structural determination relation between B and E. So this model is not correct.

5.3.3 Background Causation

Just as exciting events can cause boring ones, so too can boring events cause exciting ones. A variable C's static value at a time t, C_t, can cause a change in E's value at time t', $\Delta E_{t'}$. The gun will fire if the trigger is pulled and the safety is off. While the trigger being pulled is a productive cause of the gun's firing, the safety's being off is also a cause of the gun's firing. Had the safety not been off, the gun would not have fired. I'm going to call causal relations like these *background causal relations*. In general, a variable C's static value at t is a background cause of a variable E's change in value at t' at a world ω, according to the model $\mathcal{M}$, if, had C taken on some different value at t, E would not have changed its value as it did at t'.

(Back 1) $$\mathcal{M}, \mathcal{U}_{\omega,[f_E(t'),t']} \models \exists c \ (C_t = c \mathrel{\Box\!\!\to} \neg \Delta E_{t'})$$

As before, there are three ways that the consequent of this counterfactual could be satisfied: either E could change its value earlier or later than t', or it could change its value to something else at t', or it could fail to change its value at all. And, as before, C_t is a background cause of $\Delta E_{t'}$, at ω according to $\mathcal{M}$, if $\Delta E_{t'}$ *locally* counterfactually depends upon C_t.

(Back 2) $$\phi_E, \mathbf{PA}(E)_{\omega,[f_E(t'),t']} \models \exists c \ (C_t = c \mathrel{\Box\!\!\to} \neg \Delta E_{t'})$$

C_t is a background cause of $\Delta E_{t'}$ at ω if either (Back 1) or (Back 2) is satisfied for some structural equations model which is correct at ω.

(Back) $$\exists \mathcal{M} : \text{(Back 1) or (Back 2)} \implies \mathcal{C}(C_t, \Delta E_{t'})$$

For an example of background causation in which we have local counterfactual dependence without global counterfactual dependence, consider the neuron diagram

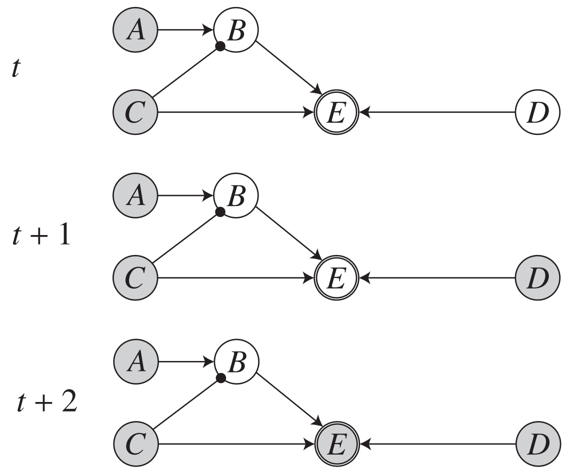


Figure 43

displayed in figure 43. In figure 43, E is a *dull* neuron, which requires two stimulations to fire. It takes 1 second for changes in the state of either C or D to affect the state of E, whereas it takes $1/2$ seconds for a change in the state of C to affect the state of B, and $1/2$ seconds for a change in the state of B to affect the state of E. At t, both A and C are firing (they continue to fire throughout the period $[t, t+2]$. At $t+1$, D fires. So, at $t+2$, E fires. Here, C's firing at $t+1$ is a background cause of E's firing at $t+2$. However, had C not been firing at $t+1$, B would have been firing at $t+3/2$, and, therefore, E still would have fired at $t+2$. However, if we hold fixed the values of E's parent variables during the period $[t, t+2]$, then, had C not been firing at $t+1$, E would not have fired at $t+2$. So C_{t+1} is a background cause of ΔE_{t+2}, according to (Back).

In standard contexts, background causes are not felicitously cited as causes in causal claims. It usually sounds odd to say that the safety's being off is what caused John's gunshot wound, or that the presence of oxygen is what caused the fire. Better to tell me that Jack pulled the trigger, or that Chloë lit a match. But that's not to say that the safety's being off didn't cause the gunshot wound or that the presence of oxygen didn't cause the fire. They both did. Indeed, there are explanatory contexts in which it is entirely appropriate to cite background causes, and inappropriate to cite productive causes. When a gun safety demonstration goes awry and John is shot, or a fire starts in a clean room which was not properly drained of oxygen, these are precisely the causes we are most apt to mention. However, in most standard contexts, it appears that we prefer to focus on productive causes. Indeed, note how intuitions shift when we alter the fire case to be an instance of productive causation. Suppose that, in the clean room, Chloë is handling highly flammable chemicals when the hapless Daniel accidentally opens the oxygen tank. A fire ensues. In this case, there is no hesitation to say that opening the oxygen tank caused the fire.

I am willing to accept that the presence of oxygen caused the match to light, and that the safety's being off caused John's gunshot wound. Does this mean that I, like Lewis, think that distinguishing between causes and background or enabling conditions is "invidious discrimination"[11]? The answer is 'no', twice over. First, my metaphysics draws a sharp distinction between productive and background causation. Both are causes, but causes of different stripes. Second, because my account requires the presence of structural determination relations between variables in order to have any causation at all, there are many events that Lewis will count as causes that I will count as mere enabling conditions. Had Caesar not been born, he wouldn't have died. So, Lewis concludes, his birth caused his death. I reject this inference. Even though, had Caesar not been born, he wouldn't have died, this is not a causal counterfactual. For it is not the case that the variable D, which takes the value 1 if Caesar dies and takes the value 0 if Caesar goes on living, is structurally determined by the variable B, which takes the value 1 if Caesar is born and 0 if he isn't. Alterations in the state of the world which set B to 0 at the appropriate time time-evolve, according to the laws of nature, into states of the world in which D is undefined. There is another variable, of course, call it 'D'', which takes the value 1 if Caesar dies and takes the value 0 otherwise. Even if this variable is natural enough to enter into causal relations, it is not the variable we are talking about when we talk about the causes of somebody's death. In normal contexts, we are interested in what caused Caesar to die, as opposed to continuing to live; and not what caused Caesar to die, as opposed to never having been born. And that is enough for my account to tell us that, in all but the most outré contexts prompted by philosophers' flights-of-fancy,[12] the sentence 'Caesar's birth caused him to die' is false. So not all discrimination is invidious, on the dynamic account.

5.3.4 Sustaining Causation

Finally, boring events can cause other boring events. A variable C's static value at a time can cause another variable E's static value at a time. The light switch's being on is causing the room to be illuminated. As I stand on the scale, my weight is causing it

[11] Lewis (1973a, p. 559)

[12] See Lewis (2004, p. 101)

to read a certain number. I'm going to call causal relations like these *sustaining causal relations*. In general, at a world ω, according to a model $\mathcal{M}$, a variable C's value at t is a sustaining cause of a variable E's value at t' if, had C taken on some different value at t, E would have taken on a different value at t'.

$$\text{(Sust 1)} \qquad \mathcal{M}, \mathcal{U}_{\omega,[f_E(t'),t']} \models \exists c \ (C_t = c \mathrel{\Box\!\!\rightarrow} E_{t'} \neq E_{\omega,t})$$

Additionally, a variable C's static value at t can be a sustaining cause of another variable's static value at t' if E's value *locally* counterfactually depends upon C's value.

$$\text{(Sust 2)} \qquad \phi_E, \mathbf{PA}(E)_{\omega,[f_E(t'),t']} \models \exists c \ (C_t = c \mathrel{\Box\!\!\rightarrow} E_{t'} \neq E_{\omega,t})$$

Electricity is flowing from the power plant, and the room is illuminated. The flow of electricity from the power plant is a sustaining cause of the room's illumination. However, had the electricity stopped flowing, the backup generator would have instantly kicked in, and the room would have been illuminated all the same. The causal counterfactual 'Had the electricity not been flowing from the power plant, the room wouldn't have been illuminated' is false. However, holding fixed the fact that the backup generator is not turned on, the *local* causal counterfactual 'Had the electricity not been flowing from the power plant, the room wouldn't have been illuminated' is true. So the room's illumination locally depends upon the electricity flowing from the power plant, even though it doesn't depend upon it globally.

Either (Sust 1) or (Sust 2) is sufficient for C_t to be a cause of $E_{t'}$.

$$\text{(Sust)} \qquad \exists \mathcal{M} : \ \text{(Sust 1) or (Sust 2)} \quad \Longrightarrow \quad \mathcal{C}(C_t, E_{t'})$$

That is, C_t is a sustaining cause of $E_{t'}$ if $E_{t'}$ counterfactually depends, either globally or locally, upon C_t in some correct structural equations model.

5.3.5 Transitivity

In all its guises, counterfactual dependence—either global or local—is sufficient for causation. However, it is not necessary. Sometimes, we trace out a sequence of events such that each event in the sequence counterfactually depends upon its predecessor in the sequence and we judge the first event to have caused the last, even though the last event does not itself counterfactually depend upon the first. When can we do this? Lewis (1973a, 2000) gave the answer 'always'. This answer allowed him to deal with cases of preemption, but it came at a cost. Chris smokes, contracts cancer, undergoes chemo, and survives. The survival counterfactually depends upon the chemo, the chemo counterfactually depends upon the cancer, and the cancer counterfactually depends upon the smoking. LEWIS concludes that the smoking caused the survival. This is difficult to swallow, no matter how it's seasoned. The answer to give is 'sometimes', and the difficulty lies in working out just when.

Recently, the idea that some distinction between *defaults* and *deviations* is central to causation has been picking up steam. MAUDLIN (2004), HITCHCOCK (2007), HALL (2007), HALPERN (2008), HITCHCOCK & KNOBE (2009), and PAUL & HALL (2013) all claim that a proper understanding of the causal relation must incorporate some standards of normality and deviations from normality. The basic idea, articulated beautifully by MAUDLIN, is that causation requires a kind of *quasi-Newtonian* system, specifying an inertial state and the kinds of forces that will kick the system out of its inertial state. The inertial state is the default; non-inertial states are deviations.

I propose that the distinction between defaults and deviations enters into the metaphysics of causation by providing the correct answer to the question of when we can trace out a chain of events such that each link in the chain counterfactually depends upon its predecessor and say that the first event in the chain caused the last, even though the last doesn't counterfactually depend upon the first. We can do so when three conditions are met: 1) the chain doesn't leave out any crucial part of the causal process leading from the first event to the last; 2) all of the links in the chain are *non-inertial*, *interesting*, or *noteworthy* events, and 3) the links in the chain are *counterfactually interlocked*. Each of these conditions requires explanation. On condition (1): what it is for the chain to not leave out any part of the causal process leading from the first event to the last is for it to be the case that the variables appearing in the chain form a directed path in a correct causal model. On condition (2): The events in the chain must all be *interesting*, *noteworthy* deviations from inertial conditions. Changes in variable values are always interesting and noteworthy deviations. However, static variable values can also be noteworthy deviations. When they are, and when there is a counterfactually interlocked chain of them lying upon a directed path in a correct structural equations model, then causes of causes will be causes. On condition (3): For each of the consequents of the counterfactuals appearing in (PROD), (PREV), (BACK), and (SUST), there are multiple ways that they may be satisfied. It may be, for instance, that in the counterfactual '$\neg \Delta C_t \;\Box\!\!\rightarrow \neg \Delta E_{t'}$', were C to have remained at its original value at t, E would still have changed at t', but it would have changed to a different value e' at t'. On the other hand, it might be that in some other counterfactual $\neg\Delta E_{t'} \Box\!\!\rightarrow \neg\Delta F_{t''}$, the antecedent which makes the consequent true is *not* that E changes to e' at t', but rather than E doesn't change at all at t'. Two counterfactuals such as these are not *interlocking*. The variable values which make the first's consequent true at not the variable values which are considered in the second's antecedent. Imagine that we have a sequence of counterfactuals $Cf_1 \rightarrow Cf_2$. If (and only if) the variable values making Cf_1's consequent true are the ones considered in evaluating Cf_2's antecedent will Cf_1 and Cf_2 be *interlocking* counterfactuals. Imagine that there is a directed path $C \rightarrow D \rightarrow E$ in some correct model, and some deviant (change in) variable value C_t causes some deviant (change in) variable value $D_{t'}$, which causes some deviant (change in) variable value $E_{t''}$, but $E_{t''}$ does not counterfactually depend upon C_t directly. The counterfactuals which must be considered to see whether C_t nevertheless causes $E_{t''}$ may not be the same counterfactuals which had to be checked to see whether C_t causes $D_{t'}$ and whether $D_{t'}$ causes $E_{t''}$—for those counterfactuals might have failed to be interlocking. This doesn't automatically mean that C_t didn't cause $E_{t''}$. We must check the interlocking counterfactuals to find out.

5.3.5.1 Tampering

For an example in which condition (3) above becomes relevant, consider the neuron diagram shown in figure 44 (from Paul & Hall, 2013, p. 96).[13] In that neuron diagram, neurons may fire with different intensities. They may either fire weakly (illustrated by the neuron being colored light gray) or strongly (illustrated by the neuron being colored dark gray). If A fires strongly, then, if C doesn't fire, then B will fire strongly, as shown in figure 44(b). If, however, A fires strongly and C fires (weakly or strongly), then B will fire weakly. E, however, is a *sensitive* neuron. It will fire strongly so long as it receives *any* signal, however strong. So, even if B fires weakly, as in figure 44(a), E will still fire strongly.

In the canonical structural equations model, each of the variables (except E) have three possible values, 0 (if their corresponding neurons don't fire at all), 1 (if they fire weakly), and 2 (if they fire strongly). The following structural equations will then be

[13] This case comes up during Paul & Hall's discussion of a case that they call 'tampering.'

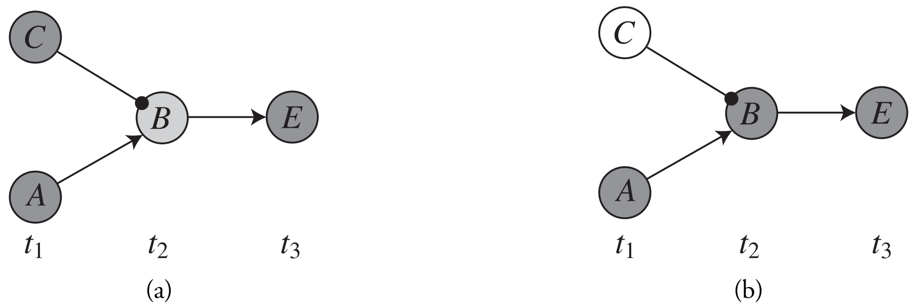


Figure 44

in force.

$$E_t := \begin{cases} 2 & \text{if } B_{t-1} \neq 0 \\ 0 & \text{if } B_{t-1} = 0 \end{cases} \qquad \forall t \in \mathbf{T}$$

$$B_t := \begin{cases} 2 & \text{if } A_{t-1} = 2 \text{ and } C_{t-1} = 0 \\ 1 & \text{if } A_{t-1} = 1 \text{ and } C_{t-1} = 0 \text{ or } A_{t-1} = 2 \text{ and } C_{t-1} \neq 0 \\ 0 & \text{if } A_{t-1} = 0 \end{cases} \qquad \forall t \in \mathbf{T}$$

Suppose that both A and C fire strongly, as shown in figure 44(a). It will then follow that, had C not fired at t_1, B would not have fired as it did at t_2.

$$\neg\Delta(0,2)^C_{t_1} \;\Box\!\!\rightarrow\; \neg\Delta(0,1)^B_{t_2} \tag{5.1}$$

That's because, had C remained at 0, B would have fired strongly, and would have changed from 0 to 2 at t_2—$\Delta(0,2)^B_{t_2}$. It will also follow that, have B not fired at t_2, E would not have fired at t_3.

$$\neg\Delta(0,1)^B_{t_2} \;\Box\!\!\rightarrow\; \neg\Delta(0,2)^E_{t_3} \tag{5.2}$$

However, we cannot conclude, on the dynamic account, that C's firing caused E's firing. The reason is that (5.1) and (5.2) are not interlocking. The variable values which make the consequent of (5.1) true when we mutilate the model so as to keep C at 0 at t_1—*viz.*, $\Delta(0,2)^B_{t_2}$—are not the variable values considered in the antecedent of (5.2). In evaluating the antecedent of (5.2), we mutilate the model so that B remains at its initial value of 0 at t_2. According to the dynamic account, however, in order to see whether C's firing caused E's firing, we must consider the counterfactual which interlocks with (5.1), namely (5.3).

$$\Delta(0,2)^B_{t_2} \;\Box\!\!\rightarrow\; \neg\Delta(0,2)^E_{t_3} \tag{5.3}$$

$$\mathcal{E}_{45} = \begin{pmatrix} M_t := F_{t-\delta} & \forall t \in \mathbf{T}_1 \\ L_t := \overline{F}_{t-\delta-\epsilon} \vee M_{t-\epsilon} & \forall t \in \mathbf{T}_2 \end{pmatrix}$$

$$\begin{tikzcd} & M \arrow[dr] & \\ F \arrow[ur] \arrow[rr] & & L \end{tikzcd}$$

Figure 45

However, this counterfactual is false. So, the dynamic account rules that C's firing did not cause E's firing. And, in this, it rules justly. While C's firing caused B to fire *as it did*, and while B's firing as it did caused E's firing, the manner in which B fired made no difference to whether E fires, so long as it fired. Therefore, since C's firing made no difference, globally or locally, to whether B fired, C's firing did not cause E's firing.

5.3.5.2 Boulder

A boulder is dislodged, and begins rolling toward Hiker. Before it reaches him, Hiker sees the boulder and moves out of the way. The boulder rolls past him, and Hiker survives.[14] The boulder's falling caused Hiker to move out of the way. Hiker's moving out of the way caused him to survive. But the boulder's falling didn't cause Hiker to survive.

We can model the structural determination relations in this case with the system of dynamic structural equations $\mathcal{E}_{45}$. In $\mathcal{E}_{45}$, F, M, and L are all a binary variables, and F is 1 at t if the boulder is *f*alling at t and is 0 if it hasn't fallen at t, M is 1 at t if Hiker has *m*oved out of the way at t and is 0 if he is underneath the boulder at t, and L is 1 if Hiker is a*l*ive at t and is 0 if Hiker is dead at t. $\mathbf{T}_1$ is the set of times during which Hiker is beneath the loose boulder. Once Hiker is clear of the loose boulder, whether he moves is no longer determined by whether or not the boulder falls.

Suppose that at t_0, the boulder falls. Seeing the boulder, Hiker moves at $t_0 + \delta$. Consequently, at $t_0 + \delta + \epsilon$, Hiker is still alive. Since, were the boulder to not fall at t_0, *i.e.*, were F_{t_0} to be 0, Hiker would not have moved at $t_0 + \delta$,

$$\neg\Delta(0,1)^{F}_{t_0} \mathrel{\Box\!\!\rightarrow} \neg\Delta(0,1)^{M}_{t_0+\delta} \tag{5.4}$$

(Prod) tells us that the boulder's falling caused Hiker's moving. And since, were Hiker not to move, he would have died—*i.e.*, were M to have stayed at 0, $L_{t_0+\delta+\epsilon}$ would

[14] This example is adapted from Hitchcock (2001), who attributes it to an early draft of Hall (2004).

have been 0,

$$\neg\Delta(0,1)^{M}_{t_0+\delta} \mathrel{\Box\!\!\rightarrow} L_{t_0+\delta+\epsilon} \neq 1 \tag{5.5}$$

(Prev) tells us that Hiker's moving caused him to continue living. However, had the boulder not fallen, Hiker would still have continued living at $t_0 + \delta + \epsilon$. So we don't have counterfactual dependence, either local or global, between the boulder's falling and Hiker's survival. We do, however, have a chain of counterfactual dependence running from the boulder's falling through Hiker's moving to Hiker's survival. F, M, and L lie on a directed path in a correct structural equations model, and the counterfactuals (5.4) and (5.5) are interlocking. Can we conclude that the boulder's falling caused Hiker's survival? Only if Hiker's survival is an interesting or noteworthy deviation. Survival is, however, Hiker's default, inertial state. So the boulder's falling did not cause Hiker's survival, on the dynamic account.

The dynamic account ends up saying precisely the same thing about Chris's survival. The smoking causes the cancer and the chemo; the cancer causes the chemo; and the chemo causes the survival. However, neither the smoking nor the cancer cause the survival. Because survival is an inertial state, we cannot take the ancestral of these interlocking counterfactuals and conclude that the smoking, or the cancer, caused the survival. We'll see other failures of transitivity below.

5.3.6 The Dynamic Account in Summary

In summary, on the dynamic account, there are four species of the causal relation, and a simple causal counterfactual conditional is sufficient for each.

$$\begin{array}{lrcl}
\textit{Productive}: & \neg\Delta C_t \mathrel{\Box\!\!\rightarrow} \neg\Delta E_{t'} & \Longrightarrow & \mathcal{C}\,(\Delta C_t, \Delta E_{t'}) \\
\textit{Preventive}: & \neg\Delta C_t \mathrel{\Box\!\!\rightarrow} E_{t'} \neq E_{\omega,t'} & \Longrightarrow & \mathcal{C}\,(\Delta C_t, E_{t'}) \\
\textit{Background}: & \exists c(C_t = c \mathrel{\Box\!\!\rightarrow} \neg\Delta E_{t'}) & \Longrightarrow & \mathcal{C}\,(C_t, \Delta E_{t'}) \\
\textit{Sustaining}: & \exists c(C_t = c \mathrel{\Box\!\!\rightarrow} E_{t'} \neq E_{\omega,t'}) & \Longrightarrow & \mathcal{C}\,(C_t, E_{t'})
\end{array}$$

In each case, the counterfactual relationship must hold in *some* correct structural equations model. The counterfactuals can be evaluated either with respect to the entire system of structural equations or with respect to E's structural equation alone—local counterfactual dependence is counterfactual dependence enough. If there is an interlocking chain of counterfactuals between deviant, interesting, or exciting events lying along a directed path in a correct structural equations model, then the first event in the chain caused the last event in the chain—whether it counterfactually depends upon it or not. The disjunction of these sufficient conditions for causation is necessary—and therefore necessary and sufficient—for causation.

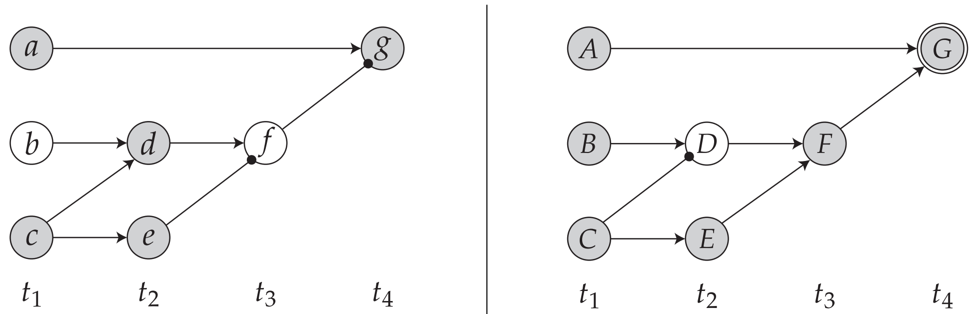
We shouldn't read too much into the labels 'productive', 'preventive', 'background', and 'sustaining'. These are evocative names, but they are used here stipulatively. I don't pretend, for instance, that every instance of preventive causation is intuitively a case of prevention.

We have to be careful in our understanding of 'ΔC_t' and 'C_t'. The value of C may in fact change at t and yet the relation still be a background or sustaining causal relation. The difference lies in the kind of counterfactual relationship between C and E. If it's true that *were the change to fail to occur*—that is, were C to remain at its initial value—E wouldn't have changed, or wouldn't have changed at t', or would have changed to a different value at t', then C is a productive cause of E. However, if C has multiple values that would have allowed E to change, and it changed from one of those values to another, then it wasn't the *change* in C which is responsible for E's change. In that case, while it's not true that, *were the change to fail to occur*, E wouldn't have changed, it may still be true that, *were C to take on some different value*, then E wouldn't have changed, or wouldn't have changed at t', or would have changed to a different value at t'. In that case, even though C's value changed at t, C's value at t was a background, rather than a productive, cause of E's change at t' (we'll see a case like this in §5.3.11 below).

5.3.7 Short-Circuits

As I mentioned in chapter 4, Hall (2007) provides a problem case for accounts of causation formulated using structural equations models. The problem is that a causal system involving 'short circuits', like the neuron diagram shown in figure 46(a), can be modeled with a system of structural equations isomorphic to a system of structural equations which models a case of preemption, such as the one shown in figure 46(b). (In figure 46(b), G is a *dull* neuron. It will only fire if it receives *two* stimulations.)

For illustration, the neuron diagram shown in figure 46(a) can be modeled with the system of structural equations shown in figure 47. (The variables are given the obvious interpretation, with the value 1 corresponding to the associated neuron firing, and the value 0 corresponding to it remaining dormant.) And the neuron diagram displayed in figure 46(b) can be modeled with the system of equations shown in figure



(a) a short circuit (b) a case of preemption

Figure 46

$$\mathcal{E}_{47} = \left(\begin{array}{lll} g_t & := a_{t-3} \wedge \overline{f}_{t-1} & \\ f_t & := d_{t-1} \wedge \overline{e}_{t-1} & \forall t \in \mathbf{T} \\ d_t & := b_{t-1} \vee c_{t-1} & \\ e_t & := c_{t-1} & \end{array} \right)$$

$$\begin{tikzcd} b \arrow[r] & d \arrow[r] & f \arrow[r] & g \\ & c \arrow[u] \arrow[r] & e \arrow[u] & a \arrow[u] \end{tikzcd}$$

Figure 47

48 (where, following the same conventions, $F^* \stackrel{\text{def}}{=} \overline{F}$, $D^* \stackrel{\text{def}}{=} \overline{D}$, and $B^* \stackrel{\text{def}}{=} \overline{B}$).

Hall (2007)'s challenge is this: not only are $\mathcal{E}_{47}$ and $\mathcal{E}_{48}$ isomorphic to one another, but additionally, the corresponding variables in each system of equations take on the same values at their respective times. So, if we are to use correct structural equations models, and a specification of the variable values, to derive information about the world's singular causal structure, then it should be that c's firing caused g's firing if and only if C's firing caused G's firing. But c's firing didn't cause g's firing, and C's firing *did* cause G's firing. So we need more information than the simple structural equations models are capable of giving us.

To see how the dynamic account handles this case, notice that, had C not fired, E would not have fired.

$$\neg\Delta(0,1)^{C}_{t_1} \mathrel{\Box\!\!\rightarrow} \neg\Delta(0,1)^{E}_{t_2} \tag{5.6}$$

And, had E not fired, F would not have fired.

$$\neg\Delta(0,1)^{E}_{t_2} \mathrel{\Box\!\!\rightarrow} \neg\Delta(1,0)^{F^*}_{t_3} \tag{5.7}$$

Finally, had F not fired, G would not have fired.

$$\neg\Delta(1,0)^{F^*}_{t_3} \mathrel{\Box\!\!\rightarrow} \Delta(0,1)^{G}_{t_4} \tag{5.8}$$

$$\mathcal{E}_{48} = \begin{pmatrix} G_t & := & A_{t-3} \wedge \overline{F^*}_{t-1} \\ F^*_t & := & D^*_{t-1} \wedge \overline{E}_{t-1} \\ D^*_t & := & B^*_{t-1} \vee C_{t-1} \\ E_t & := & C_{t-1} \end{pmatrix}$$

$$\begin{tikzcd} B^* \arrow[r] & D^* \arrow[r] & F^* \arrow[r] & G \\ & C \arrow[u] \arrow[r] & E \arrow[u] & A \arrow[u] \end{tikzcd}$$

Figure 48

Each of these events are changes in variable values, so they are all interesting deviations, $C \to E \to F^* \to G$ is a directed path, and (5.6–5.8) are interlocking counterfactuals, so it follows on the dynamic account that $\Delta(0,1)^C_{t_1}$ is a productive cause of $\Delta(0,1)^G_{t_4}$.

Parallel reasoning does not apply to the short circuit. While there is a sequence of interlocking counterfactual dependence running from c's firing to g's firing which lies on a directed path, namely

$$\Delta(0,1)^c_{t_1} \to \Delta(0,1)^e_{t_2} \to 0^f_{t_3} \to \Delta(0,1)^g_{t_4}$$

$0^f_{t_3}$ is a boring, default event. Remaining dormant is the inertial state of a neuron. Nor does g's firing counterfactually depend, either globally or locally, on c's firing in any correct structural equations model. So we cannot conclude that c's firing caused g's firing.

Note that, without condition (1) from §5.3.5—*i.e.*, the condition that the interlocking counterfactuals relate events whose variables lie along a directed path—we *would* be able to conclude that $\mathcal{C}\ (\Delta c_{t_1}, \Delta g_{t_4})$, since the counterfactuals

$$\neg \Delta c_{t_1} \;\Box\!\!\rightarrow \neg \Delta e_{t_2}$$
$$\neg \Delta e_{t_2} \;\Box\!\!\rightarrow \neg \Delta g_{t_4}$$

are both true, interlocking, and relate interesting, deviant events.

5.3.8 Switching

A train approaches a fork in the tracks. At the fork, the switch leads off to the left. So, when the train arrives at the fork, it travels off on the left-hand track, rather than the right. Since the tracks reconverge up ahead, the train arrives at its destination just as it would if it had taken the right-hand track.[15]

[15] Adapted from Hall (2000, p. 205)

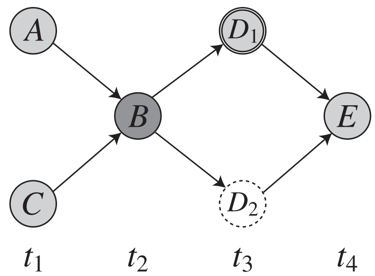


Figure 49: A switching case involving a chain of productive causation.

Here, the argument goes, the switch being set to the left caused the train to take the left-hand track; and it's taking the left-hand track caused it to arrive at the station. But the switch being set to the left didn't cause the train to arrive at the station.[16]

In a natural structural equations model, it follows that, had the switch been set to the right, the train wouldn't have taken the left-hand track. So, the switch's being set to the left is a background cause of the train's taking the left-hand track. It's less clear in what sense, had the train not taken the left-hand-track, it would not have arrived at the station. However, we needn't fuss over this. Since the switch being set to the left is not a deviant, interesting, or exciting event—being set to the left, like being set to the right, is an inertial state for the switch—we don't get to trace out a causal path from it to the train's arrival at the station. So the dynamic account doesn't conclude that the switch's being set to the left caused the train to arrive.

However, not every case that's been called a case of switching in the literature is of this form. Sometimes, cases involving chains of productive causation are presented as cases of switching. A representative example is displayed in the neuron diagram in figure 49.[17] In figure 49, B is a neuron which can fire either weakly or strongly. If it receives one stimulation, it will fire weakly. If it receives two, then it will fire strongly (strong firing is represented with a darker shade of gray). D_1 is a dull neuron which will only fire if it receives a strong stimulation. D_2 is a fragile neuron, which will fire if it receives a weak stimulation, but which will burst, and not fire, if it receives a strong stimulation. Then, if C doesn't fire, the signal from A will travel along the lower path via D_2. If C does fire, then the signal from A will travel along the upper path via D_1. Either way, E will fire. Now, if C hadn't fired, then B wouldn't have

[16] I'm a bit uneasy with the second causal relation in this purported counterexample, but it's not my focus now, so I'll let it be. There are other cases with the same structure in which the second link in the causal chain is far less objectionable.

[17] Similar neuron diagrams can be found in Hitchcock (2009) and Paul & Hall (2013, p. 232)

fired strongly (it would have fired weakly); and, had B fired weakly, then D_1 would not have fired; moreover, had D_1 not fired, E would not have fired. So, there is a chain of interlocking counterfactual dependence lying along a directed path leading from C's firing to E's firing. The dynamic account rules that C's firing caused E's firing.

If you are inclined to think that this is the wrong verdict, it might be because of the way the case was described. I said that "the signal from A" will either travel along the upper or lower path, and that C's firing will determine which. Language like this encourages you to think of A at the thing providing the signal, and C as the switch which determines the path it takes. But there's no good reason for this differential treatment. A and C are perfectly symmetrical in this neuron diagram. I might just as well have said, then, that, if A fires, then the signal from C will take the upper path; and if A doesn't fire, then the signal from C will take the lower path.

Because A and C are perfectly symmetrical, A's firing caused E's firing iff C's firing did. We might want to say that neither A's firing nor C's firing caused E's firing. This isn't crazy—we could still say that their disjunction caused E's firing. This is, indeed, exactly what the dynamic account says about simple symmetrical overdetermination cases (see §5.3.9). But neither is there anything crazy about saying that *both* A's and C's firing caused E's firing. Like symmetric overdetermination cases, this is a case about which intuition delivers no clear verdict—so long as we're careful to not think about A's and C's firings differently, one as signal, the other as switch. It is, therefore, a case to be settled by theory, not intuition—spoils to the victor.

5.3.9 Symmetric Overdetermination

Suppose that Suzy and Billy stand on opposite sides of the window, both throw their rocks, and both rocks strike the window at the same time. Either rock would have been sufficient, on its own, to shatter the window. However, given that Billy threw, the window's shattering does not depend upon Suzy's throw. And, given that Suzy threw, the window's shattering does not depend upon Billy's throw. Did Suzy's throw cause the window to shatter? Intuition is split. There is some inclination to say that it did, since, obviously, either Billy's or Suzy's throw caused it, and there's nothing that Billy's throw has that Suzy's doesn't. So, if either of them caused it, then both of them did. And it's not the case that *neither* of their throws caused it, since then, it seems, the window's shattering would be uncaused. We should resist this seductive reasoning. Just because neither Suzy's nor Billy's throw individually caused the window to shatter, this doesn't mean that the shattering is uncaused, since it can be caused by the *disjunction* of Suzy's and Billy's throws without being caused by either throw individually.[18]

5.3.9.1 Disjunctive Causes

If we countenance disjunctions of (changes in) variable values as causal relata, then the simple counterfactual condition provides a sufficient condition for these disjunctions causing (a change in) another variable's value. For instance, the simple counterfactual condition tells us that $\Delta C_{t_1} \vee \Delta C'_{t_2}$ caused ΔE_{t_3} if, had both C and C' stayed at their original values at t_1 and t_2, respectively, then E would either have not changed, or not have changed at t_3, or have changed to a different value at t_3.

$$(\neg\Delta C_{t_1} \wedge \neg\Delta C'_{t_2}) \mathrel{\Box\!\!\rightarrow} \neg\Delta E_{t_3}$$

It is straightforward to evaluate causal counterfactuals like these, given a correct dynamic structural equations model. It simply involves one extra intervention. As before, local counterfactual dependence is dependence enough.

When causes are disjunctive, however, we must emend the account of transitivity that I offered back in §5.3.1. Call a counterfactual conditional stating that if one deviant event hadn't occurred, another deviant event wouldn't have occurred a 'deviant counterfactual.' Then, the new wrinkle is this: if there is a chain of interlocking deviant counterfactuals running from C to D along a directed path, and another chain of interlocking deviant counterfactuals running from C to D' along a directed path, $D, D' \in \mathbf{PA}(E)$, and there is a true, interlocking counterfactual of the form '$\neg(D \vee D') \mathrel{\Box\!\!\rightarrow} \neg E$', then C caused E.[19]

Consider, for instance, the neuron diagram on display in figure 50. There, C's firing caused E's firing. However, the counterfactual conditional $\neg\Delta C_{t_1} \mathrel{\Box\!\!\rightarrow} \neg\Delta E_{t_4}$ is false. Nor is there any correct model within which C is a parent of E, since this would involve a violation of autonomy. However, the following interlocking deviant (local) counterfactuals are true.

$$\neg\Delta C_{t_1} \mathrel{\Box\!\!\rightarrow} \neg\Delta D_{1t_3}$$
$$\neg\Delta C_{t_1} \mathrel{\Box\!\!\rightarrow} \neg\Delta D_{2t_3}$$

[18] Sartorio (2006) offers a defense of disjunctive causes.

[19] Here, I'm using 'C' and the like to refer to an event involving the value of the variable C, whether C's value changes or not.

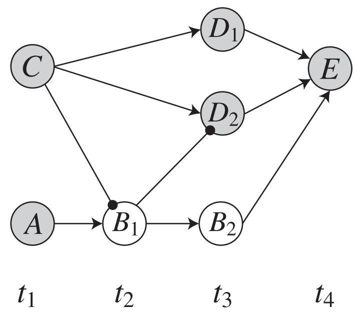


Figure 50

and, moreover,

$$(\neg \Delta D_{1t_3} \wedge \neg \Delta D_{2t_3}) \mathrel{\Box\!\!\rightarrow} \neg \Delta E_{t_4}$$

So C's firing caused E's firing.

This emendation could, and should, be generalized to disjunctive causes involving more than two disjuncts, but in the interests of space, I'll leave that generalization as an exercise for the reader.

5.3.9.2 Conjunctive Causes

We can similarly use the simple counterfactual test to say something about when *conjunctive* events are causes and effects. The conjunctive event $\Delta C_{t_1} \wedge \Delta C'_{t_2}$ is a cause of ΔE_{t_3} if, had either C not changed its value at t_1 or else had C' not changed its value at t_2, then E would not have changed its value at t_3.

$$(\neg \Delta C_{t_1} \vee \neg \Delta C'_{t_2}) \mathrel{\Box\!\!\rightarrow} \neg \Delta E_{t_3}$$

Here's how we evaluate this counterfactual: it is true iff every way of intervening on the actual values of C and C' so as to make the antecedent true makes the consequent true as well.[20] So, it is true iff all of the following three counterfactuals are true:

$$\begin{aligned} \neg \Delta C_{t_1} &\mathrel{\Box\!\!\rightarrow} \neg \Delta E_{t_3} \\ \neg \Delta C'_{t_2} &\mathrel{\Box\!\!\rightarrow} \neg \Delta E_{t_3} \\ (\neg \Delta C_{t_1} \wedge \neg \Delta C'_{t_2}) &\mathrel{\Box\!\!\rightarrow} \neg \Delta E_{t_3} \end{aligned}$$

The first two counterfactuals guarantee that, if $\Delta C_{t_1} \wedge \Delta C'_{t_2}$ is a cause of ΔE_{t_3}, then both ΔC_{t_1} and $\Delta C'_{t_2}$ are individually causes of ΔE_{t_3}. However, the third counterfac-

[20] This is in line with Briggs (2012)'s treatment of disjunctive antecedents.

$$\mathcal{E}_{51} = (V_{t+\delta} := \phi_V(U_t, W_t, T_t))$$

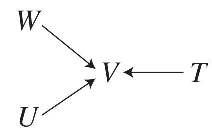


Figure 51

tual shows us that both ΔC_{t_1} and $\Delta C'_{t_2}$ individually causing ΔE_{t_3} is not sufficient for their conjunction causing ΔE_{t_3}.[21]

For an example of a case in which $\mathcal{C}(\Delta C_{t_1}, \Delta E_{t_3})$ and $\mathcal{C}\left(\Delta C'_{t_2}, \Delta E_{t_3}\right)$, yet it is not the case that $\mathcal{C}\left(\Delta C_{t_1} \wedge \Delta C'_{t_2}, \Delta E_{t_3}\right)$, consider the structural equations model shown in figure 51. In that model, V, U, W, and T are all binary variables and

$$\phi_V(U_t, W_t, T_t) = \begin{cases} 1 & \text{if } T_t = 1 \\ U_t \equiv W_t & \text{if } T_t = 0 \end{cases}$$

($* \equiv \circ$ is the familiar truth function $1 - | * - \circ |$.) That is, if $T_t = 1$, then $V_{t+\delta} = 1$. If, however, $T_t = 0$, then $V_{t+\delta} = 1$ iff U_t and W_t have the same value. Now, suppose that, at t^*, $\Delta(1,0)^T$, $\Delta(0,1)^U$, and $\Delta(1,0)^W$. Then, at $t^* + \delta$, V will change from 1 to 0. This change in V's value will be caused by $\Delta(0,1)^U_{t^*}$ since, given that W changed to 0, had U stayed at 0, V would have remained at 1. It will also be caused by $\Delta(1,0)^W_{t^*}$ since, given that U changed to 1, had W remained at 1, V would have remained at 1. However, the change in V's value will not be caused by $\Delta U_{t^*} \wedge \Delta W_{t^*}$ since, had neither U nor W changed, V would still have changed.

This account could, and should, be generalized to conjunctive causes involving more than two conjuncts, as well as cases involving preventive, background, and sustaining causation. However, in the interest of space, I will leave this, too, as an exercise for the reader.

5.3.10 Preemptive Double Prevention

Consider the neuron diagram shown in figure 52. This is a case of preemptive double prevention.[22] F is prevented from firing; and, had F fired, it would have prevented G from firing. Stopping short the signal from B prevents F's firing from preventing G's firing—double prevention. C stops short the signal from B which,

[21] In general, this means that $\mathcal{C}(C \wedge C', E)$ entails that $\mathcal{C}(C \vee C', E)$, but not *vice versa*.

[22] Throughout Collins et al. (2004), cases with this structure are simply called cases of 'preventive prevention', or 'double prevention'. However, I think it is important to distinguish this case from the neuron diagram in figure 40, as well as double prevention cases not involving any preemption.

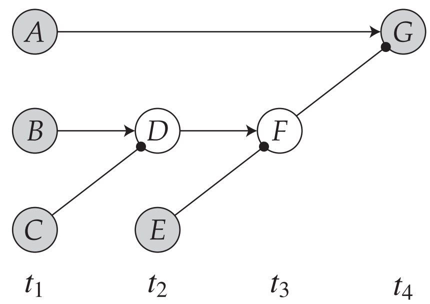


Figure 52

unabated, would have caused F to fire. However, had C not stopped this signal, E would have stopped it—preemption.

The dynamic account delivers the verdict that C's firing did not cause G's firing. However, it is not difficult to reason yourself to the opposite verdict. Either C's firing or E's firing prevented F from firing, and thereby allowed G to fire. And it wasn't E's firing that did it, so it must have been C's firing that did it. So C's firing must have caused G's firing. This reasoning is seductive, but it is a temptation to be resisted.

In this case, we judge that either C's firing or E's firing caused G's firing.

$$\mathcal{C}\ (\Delta C_{t_1} \vee \Delta E_{t_2}, \Delta G_{t_4})$$

We also judge that it is not the case that E's firing caused G's firing.

$$\neg\mathcal{C}\ (\Delta E_{t_2}, \Delta G_{t_4})$$

And we are led to conclude, via the seductive reasoning, is that it was C's firing which caused G's firing,

$$\mathcal{C}\ (\Delta C_{t_1}, \Delta G_{t_4})$$

I don't deny either of the intuitions which make up the premises of this argument. (Though the first requires some care to adequately capture. We judge that it was either C's firing or E's firing that caused G's firing. This judgment doesn't discern between $\mathcal{C}(\Delta C_{t_1}, \Delta G_{t_4}) \vee \mathcal{C}(\Delta E_{t_2}, \Delta G_{t_4})$ and $\mathcal{C}\ (\Delta C_{t_1} \vee \Delta E_{t_2}, \Delta G_{t_4})$. Therefore, an account which secures either does an adequate job capturing the judgment.) I deny the validity of the inference. If the reasoning works here, then it works in the symmetric overdetermination case as well. If we're comfortable saying that the disjunction of Suzy's and Billy's throws caused the window to shatter even though neither Billy's nor

Suzy's throw on their own caused the window to shatter, then we should also be comfortable saying that the disjunction of C's firing and E's firing caused G's firing, even though neither C's firing nor E's firing on their own caused caused G's firing. Notice that this doesn't prevent us from saying that C's firing cut short the signal from B and that E's firing did not. For C's firing prevented D's firing; but it's not the case that E's firing prevented F's firing.

5.3.11 Trumping

The soldiers will advance iff the order of the highest-ranked officer tells them to advance.[23] Let S take on the value 1 if the Sergeant orders to advance, 2 if he orders to stay put and 0 if the Sergeant issues no order. M is 1 if the Major orders to advance, 2 if he orders to stay put, and 0 if he issues no order. A is 1 if the soldiers advance and 0 if they stay put. Then, the following structural equation will be correct

$$A_t := \phi_A(M_{t-\delta}, S_{t-\delta}) \qquad\qquad M \longrightarrow A \longleftarrow S$$

where

$$\phi_A(M_t, S_t) = \begin{cases} 0 & \text{if } M_t = 2 \\ 1 & \text{if } M_t = 1 \\ 0 & \text{if } M_t = 0 \wedge S_t = 2, 0 \\ 1 & \text{if } M_t = 0 \wedge S_t = 1 \end{cases}$$

If the Major orders the soldiers to stay put, then stay put they will, regardless of what the Sergeant orders. If the Major orders the soldiers to advance, then advance they will, regardless of what the Sergeant orders. If the Major issues no order, then the soldiers will follow the orders of the Sergeant, if such there be. If the soldiers receive no orders, then they will stay put.

At t, both the Sergeant and the Major order the soldiers to advance, and advance they do. Had the Major not issued the order to advance—*i.e.*, had he remained silent at t—then the soldiers would still have advanced, since, if the Major had remained silent, the Sergeant's orders would have held sway. Similarly, had the Sergeant not issued the order to advance—*i.e.*, had he remained silent at t—then the soldiers would still have advanced, since the Major would still have issued his order. So, on the dynamic account, neither the Major nor the Sergeant productively caused the soldiers to advance. However, had *neither* the Major *nor* the Sergeant issued their orders, then

[23] See Schaffer (2004), who attributes the example to Bas van Fraassen.

the soldiers would have stayed put. So the dynamic account rules that the *disjunction* of the Major's and the Sergeant's orders is a productive cause of the soldier's advancing.

So far, the dynamic account treats this case just like a case of symmetric overdetemination. But it does not feel like a case of symmetric overdetermination. Given that the Major issued an order, it just doesn't matter what order the Sergeant gives. In this case, it appears as though the Major's higher rank breaks the tie between the two orders. The Major's order *trumps* the Sergeant's order. That is, it seems correct to say that it was the Major that caused the soldiers to advance, and it seems incorrect to say that it was the Sergeant that caused them to advance.

In fact the dynamic account secures just this result, since, even though the Major's orders do not count as a *productive* cause of the soldier's advancing, they do count as a *background* cause of the soldier's advancing. Had the Major ordered the soldiers to stay put, then they would have stayed put, even though the Sergeant had ordered them to advance. The Sergeant's orders, on the other hand, do not count as a background cause of the soldier's advancing. Given that the Major had ordered them to advance, there is nothing that the Sergeant could have done to prevent the soldiers from advancing. In my view, this captures the 'tie breaking' intuition of the case fairly well.

Note that we don't get this result if we consider the binary variables M' and S', which either take the value 0, if no order is issued, or 1, if the order to advance is issued. M''s determination of A and S''s determination of A are perfectly symmetrical. These structural determination relations are representable with the equation $A_t := M'_{t-\delta} \vee S'_{t-\delta}$. So if both M' and S' change from 0 to 1 at t, $\Delta M'_t$ caused $\Delta A_{t+\delta}$ iff $\Delta S'_t$ caused $\Delta A_{t+\delta}$, and M'_t caused $\Delta A_{t+\delta}$ iff S'_t caused $\Delta A_{t+\delta}$. Due to the perfect symmetry, there is no sense in which M' could trump S'. Insofar as we are thinking about the property of the world described by *the Major's orders* in such a way that they could trump the Sergeant's orders, this shows that we are thinking about a property of the world better represented by M than by M'.[24]

5.4 In Summation

5.4.1 Looking Back

The foregoing three chapters have outlined an account of singular causation according to which causes are, in some good sense, both nomically sufficient and nom-

[24] Similar diagnoses of the trumping case are provided by Lewis (2000) and Halpern & Hitchcock (2010).

ically necessary for their effects. The sense in which they are nomically sufficient is that, in order for the event of a certain part of the world having or acquiring a certain property (*i.e.*, the event of a variable taking on or retaining a certain value) to be causally related to the event of some other part of the world having or acquiring a certain property (*i.e.*, the event of another variable taking on or retaining a certain value), it is necessary that the first variable enters into a network of *structural determination* with the second. What it is for a variable to enter into a network of structural determination like this is for its value (perhaps in tandem with other variables' values) to be nomically sufficient for the value of its structural descendants throughout a certain swath of modal space.

Because structural determination relations are, in some sense, more sparse between high-level variables than they are between low-level variables (see figure 2), this account is causally emergentist, in the sense of chapter 2. The structural determination relations found at the low-level, fundamental physical level of description are not the only structural determination relations that there are. There are additionally structural determination relations between high-level variables which do not reduce to, and do not even supervene upon, the structural determination relations between the low-level variables whose values their values supervene upon. Therefore, given that these determination relations are necessary for singular causation—they provide the pathways along which singular causal influence propagates—causal relations between high-level events do not reduce to causal relations between low-level events. In providing an emergentist account such as this, there is the danger that, by soaring to such ontic heights, we will lose sight of more minute details which end up making a difference to the question of which events singularly caused which other events. This led us into the problems of chapter 4, where, by ignoring too much of the world's causal structure, we lost the ability to correctly discriminate causes from non-causes. There, I suggested that the solution lies in maintaining that a network of structural determination is correctly represented by a structural equations model only if each of the determination relations in our model are *autonomous* of each other.

In this chapter, I've offered an account of the sense in which causes are nomically necessary for their effects by outlining a counterfactual account of singular causation, dubbed 'the dynamic counterfactual account'. This account is broadly in the spirit of Lewis's original 1973 counterfactual account. The primary differences are four-fold: firstly, whereas Lewis's counterfactuals concerned the occurrence or nonoccurrence of *events*, the dynamic account simply looks at the counterfactual relationship between (changes in) variable values at times. It may therefore remain neutral on questions about the modal profile of the corresponding events. Secondly, whereas Lewis evaluated his counterfactuals using the 'closest possible world' analysis, the dynamic account evaluates its counterfactuals relative to a network of structural determination between variable values. Thirdly, whereas, for Lewis, it did not matter what *kinds* of events were being related—neither whether they were events in which something *changed* or whether something remained the same, nor whether they were events in which something noteworthy or deviant happened or whether something boring and default happened—for the dynamic account, these differences matter. Fourthly and relatedly, whereas Lewis thought that, *whenever* there was a chain of counterfactual dependence leading from C to E, C caused E, according to the dynamic account, such chains of counterfactual dependence only entail that C caused E when three additional conditions are met: 1) the chain doesn't leave out any important part of the causal process leading from C to E, 2) every event along the chain is an interesting or noteworthy deviation from normality, and 3) the counterfactuals in the chain are all *interlocking*.

I've shown that these emendations to Lewis's original counterfactual account are able to deal with many of the most pressing counterexamples which that account faced. Indeed, I believe that the account deals with many other counterexamples to the counterfactual analysis as well,[25] though a defense of that claim must be postponed. The dynamic account should, of course, be generalized to deal with cases of indeterministic causation, as well as possible cases of backwards causation. Though I won't be providing those generalizations here, I believe that both problems are dealt with by paying more attention to the structural dependence relations that underlie the causal counterfactuals appearing in the dynamic account. In particular, I believe that, when backwards causation is possible, the past state of the world is structurally determined by the future state of the world. And, when causation is indeterministic, it is because the underlying structural determination relations are indeterministic.

[25] To name but a few: Dretske (1977), McDermott (1995), Hitchcock (1996), Thomson (2003), Strevens (2003), Collins (2004), McGrath (2005), Schaffer (2005), Hitchcock & Knobe (2009), Schaffer (2012), Ballarin (forthcoming), and every counterexample from the compendium of counterexamples provided by Paul & Hall (2013).

BIBLIOGRAPHY

Ballarin, Roberta. forthcoming. "Disjunctive Effects and the Logic of Causation." *The British Journal for the Philosophy of Science*. [169]

Baumgartner, Michael. 2013. "A Regularity Theoretic Approach to Actual Causation." *Erkenntnis*. [61]

Beebee, Helen, Christopher Hitchcock & Peter Menzies, editors. 2009. *The Oxford Handbook of Causation*. Oxford University Press, Oxford. [177]

Bennett, Jonathan. 2003. *A Philosophical Guide to Conditionals*. Clarendon Press, Oxford. [90]

Blanchard, Thomas & Jonathan Schaffer. forthcoming. "Cause without Default." In *Making a Difference*, Helen Beebee, Christopher Hitchcock & Huw Price, editors. Oxford University Press, Oxford. [125]

Briggs, Rachael. 2012. "Interventionist Counterfactuals." *Philosophical Studies*, vol. 160: 139–166. [61], [78], [163]

Cartwright, Nancy. 2009. "How to Do Things with Causes." apa *Proceedings and Addresses*, vol. 83 (2). [68]

Chalmers, David J. 2006a. "The Foundations of Two-Dimensional Semantics." In *Two-Dimensional Semantics: Foundations and Applications*, M. Garcia-Carpintero & J. Macia, editors. Oxford University Press, Oxford. [21]

—. 2006b. "Two-Dimensional Semantics." In *Oxford Handbook of the Philosophy of Language*, E. Lepore & B. Smith, editors. Oxford University Press, Oxford. [21]

—. 2012. *Constructing the World*. Oxford University Press, Oxford. [22]

Collins, John. 2004. "Preemptive Prevention." In Collins et al. (2004), chap. 4, 107–117. [113], [147], [169]

Collins, John, Ned Hall & L. A. Paul, editors. 2004. *Causation and Counterfactuals*. The MIT Press, Cambridge, ma. [164], [171], [172], [175], [177], [178]

Davidson, Donald. 1967. "Causal Relations." *The Journal of Philosophy*, vol. 64 (21): 691–703. Reprinted in *Essays on Actions and Events* (2001), Oxford University Press, 2nd edition, pp. 149–162. Page numbers are from *Essays on Actions and Events*. [38], [46], [47], [54], [55], [56], [57]

DeRose, Keith. 1999. "Can It Be That It Would Have Been Even Though It Might Not Have Been?" *Philosophical Perspectives*, vol. 13: 385–413. [95]

Dowe, Phil. 2000. *Physical Causation*. Cambridge University Press, Cambridge. [vi], [1], [2], [3], [4], [5], [7], [8], [9], [10], [11], [12], [49], [50], [58]

Dretske, Fred I. 1977. "Referring to Events." *Midwest Studies in Philosophy*, vol. 2 (1): 90–99. [169]

Earman, John. 1986. *A Primer on Determinism*. Reidel, Dordrecht. [9]

Eiter, Thomas & Thomas Lukasiewicz. 2002. "Complexity Results for Structure-Based Causality." *Artificial Intelligence*, vol. 142 (1): 53–89. [106]

Fair, David. 1979. "Causation and the Flow of Energy." *Erkenntnis*, vol. 14: 219–50. [10], [11]

Field, Hartry. 2003. "Causation in a Physical World." In *Oxford Handbook of Metaphysics*, Michael Loux & Dean Zimmerman, editors, 435–60. Oxford University Press, Oxford. [46]

Gettier, Edmund. 1963. "Is Justified True Belief Knowledge?" *Analysis*, vol. 23: 121–123. [18]

Gibbard, Allan. 2003. *Thinking How to Live*. Harvard University Press, Cambridge, MA. [16]

Glymour, Clark, David Danks, Bruce Glymour, Frederick Eberhardt, Joseph Ramsey, Richard Scheines, Peter Spirtes, Choh Man Teng & Jiji Zhang. 2010. "Actual Causation: A Stone Soup Essay." *Synthese*, vol. 175: 169–192. [104]

Glynn, Luke. 2013. "Of Miracles and Interventions." *Erkenntnis*. DOI 10.1007/s10670-013-9436-5. [61], [62], [70]

Hájek, Alan. ms. "Most Counterfactuals are False." Available at http://philrsss.anu.edu.au/people-defaults/alanh/papers/MCF.pdf. [95]

Hall, Ned. 2000. "Causation and the Price of Transitivity." *Journal of Philosophy*, vol. 97 (4): 198–222. Reprinted in Collins et al. (2004). [159]

—. 2004. "Two Concepts of Causation." In Collins et al. (2004), 225–276. [16], [155]

—. 2007. "Structural Equations and Causation." *Philosophical Studies*, vol. 132 (1): 109–136. [viii], [62], [70], [71], [104], [105], [109], [110], [111], [123], [134], [152], [157], [158]

Halpern, Joseph Y. 2008. "Defaults and Normality in Causal Structures." *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning*, 198–208. [103], [105], [106], [107], [108], [111], [112], [115], [152]

—. ms. "Appropriate Causal Models and Stability of Causation." [122], [124], [127], [128], [129]

Halpern, Joseph Y. & Christopher Hitchcock. 2010. "Actual Causation and the Art of Modeling." In *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, Rina Dechter, Hechtor Geffner & Joseph Y. Halpern, editors, 383–406. College Publications. Available at http://arxiv.org/pdf/1106.2652v1.pdf. [62], [70], [103], [104], [107], [108], [120], [121], [122], [123], [126], [129], [167]

Halpern, Joseph Y. & Judea Pearl. 2001. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." In *Proceedings of the Seventeeth Conference on Uncertainty in Artificial Intelligence*, John Breese & Daphne Koller, editors, 194–202. Morgan Kaufman, San Francisco, CA. [61], [103], [106], [107], [108], [124]

—. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." *The British Journal for the Philosophy of Science*, vol. 56: 843–887. [61], [65], [103], [105], [106], [107], [108], [122], [123], [124]

Handfield, Toby, Charles R. Twardy, Kevin B. Korb & Graham Oppy. 2008. "The Metaphysics of Causal Models: Where's the Biff?" *Erkenntnis*, vol. 68: 149–168. [13], [61]

Hausman, Daniel M. & James Woodward. 1999. "Independence, Invariance, and the Causal Markov Condition." *British Journal for the Philosophy of Science*, vol. 50: 521–583. [62], [67]

Hawthorne, John. 2005. "Chance and Counterfactuals." *Philosophy and Phenomenological Research*, vol. 70 (2): 396–405. [95]

Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. Free Press, New York.

Hiddleston, Eric. 2005a. "A Causal Theory of Counterfactuals." *Noûs*, vol. 39 (4): 632–657. [61]

—. 2005b. "Causal Powers." *The British Journal for the Philosophy of Science*, vol. 56: 27–59. [103], [104]

Hitchcock, Christopher. 1996. "The Role of Contrast in Causal and Explanatory Claims." *Synthese*, vol. 107 (3): 395–419. [169]

—. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy*, vol. 98 (6): 273–299. [61], [62], [70], [103], [125], [155]

—. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review*, vol. 116 (4): 495–532. [62], [70], [103], [104], [105], [111], [152]

—. 2009. "Structural Equations and Causation: Six Counterexamples." *Philosophical Studies*, vol. 144: 391–401. [160]

Hitchcock, Christopher & Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy*, vol. 106 (11): 587–612. [105], [111], [152], [169] Hume, David. 1975. *Enquiry concerning Human Understanding*. Clarendon Press, Oxford, third edn. [60]

Kahneman, Daniel & Dale T. Miller. 1986. "Norm Theory: Comparing Reality to Its Alternatives." *Psychological Review*, vol. 94 (2): 136–153. [111]

Kim, Jaegwon. 1973. "Causes and Counterfactuals." *Journal of Philosophy*, vol. 70 (17): 570–572. url: http://www.jstor.org/stable/2025312. [62]

—. 1976. "Events as Property Exemplifications." In *Action Theory*, Myles Brand & Douglas Walton, editors, 159–77. D. Reidel, Dordrecht. Republished in Varzi & Casati (1996, p. 117–136). [31]

—. 1984a. "Epiphenomenal and Supervenient Causation." *Midwest Studies in Philosophy*, vol. 9 (1): 257–270. [31], [32]

—. 1984b. "Supervenience and Supervenient Causation." *The Southern Journal of Philosophy*, vol. 22 (S1): 45–56. [31], [32], [33], [35], [37], [45]

Kment, Boris. 2006. "Counterfactuals and Explanation." *Mind*, vol. 115: 261–309. [87], [90]

Lewis, David K. 1970. "How to Define Theoetical Terms." *The Journal of Philosophy*, vol. 67 (13): 427–446. [8]

—. 1972. "Psychophysical and Theoretical Identifications." *Australasian Journal of Philosophy*, vol. 50 (3): 249–258. [8]

—. 1973a. "Causation." *The Journal of Philosophy*, vol. 70 (17): 556–567. [vii], [ix], [x], [1], [28], [29], [30], [40], [41], [43], [45], [58], [60], [62], [97], [143], [150], [151], [168], [169]

—. 1973b. *Counterfactuals*. Blackwell Publishers, Malden, ma. [viii], [29], [40], [63], [71], [72], [144]

—. 1979a. "Attitudes De Dicto and De Se." *The Philosophical Review*, vol. 88 (4): 513–543. [15]

—. 1979b. "Counterfactual Dependence and Time's Arrow." *Noûs*, vol. 13 (4): 455–476. [ix], [29], [40], [62], [72], [88], [95], [96]

—. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy*, vol. 61 (4): 343–377. [40]

—. 1986a. "Causation." In *Philosophical Papers*, vol. II. Oxford University Press, New York. [29], [40], [41], [109], [116], [122], [143]

—. 1986b. "Events." In *Philosophical Papers*, vol. II, 241–269. Oxford University Press, New York. [40], [41], [62], [104]

—. 1994a. "Humean Supervenience Debugged." *Mind*, vol. 103 (412): 473–490. [40]

—. 1994b. "Reduction of Mind." In *A Companion to Philosophy of Mind*, Samuel Guttenplan, editor, 412–431. Blackwell Publishers, Oxford. [10], [12]

—. 2000. "Causation as Influence." *The Journal of Philosophy*, vol. 97 (4): 182–197. Reprinted in Collins et al. (2004, pp. 75–106). [vii], [29], [30], [43], [44], [45], [51], [52], [53], [54], [57], [143], [151], [152], [167]

—. 2004. "Causation as Influence." In Collins et al. (2004), chap. 3, 75–106. [52], [150]

Lucas, Robert E. 1976. "Economic Policy Evaluation: A Critique." In *The Philips Curve and Labor Markets*, Karl Brunner & Allan H. Meltzer, editors, vol. 1 of *Carnegie-Rochester Conference Series on Public Policy*, 19–46. American Elsevier, New York. Available online at `http://www.eco.unrc.edu.ar/wp-content/uploads/2010/03/lucas1976.pdf`. [68]

Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly*, vol. 2 (4): 245–55. [49], [58], [80], [143]

Malament, David B. 2008. "Norton's Slippery Slope." *Philosophy of Science*, vol. 75: 799–816. [9]

Margolis, Eric & Stephen Laurence. 1999. "Concepts and Cognitive Science." In *Concepts: Core Readings*, Eric Margolis & Stephen Laurence, editors, chap. 1, 3–81. MIT Press, Cambridge, MA. [14]

Maudlin, Tim. 2004. "Causation, Counterfactuals, and the Third Factor." In Collins et al. (2004), 419–443. [111], [152]

—. 2007. "A Modest Proposal Concerning Laws, Counterfactuals, and Explanations." In *The Metaphysics within Physics*, 5–49. Oxford University Press, Oxford. [71], [88], [89], [90], [91], [92]

—. 2011. *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*, chap. 5: Causation, 114–147. Wiley-Blackwell, Malden, MA, third edn. [48]

McDermott, Michael. 1995. "Redundant Causation." *The British Journal for the Philosophy of Science*, vol. 46 (4): 523–544. [113], [147], [169]

McGee, Vann. 1985. "A Counterexample to Modus Ponens." *The Journal of Philosophy*, vol. 82 (9): 462–471. [78]

McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." *Philosophical Studies*, vol. 123: 125–148. [169]

McLaughlin, Brian P. 1992. "The Rise and Fall of British Emergentism." In *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, Ansgar Beckermann, Hans Flohr & Jaegwon Kim, editors, 49–93. Walter de Gruyter, Berlin. [39]

Menzies, Peter. 1988. "Against Causal Reductionism." *Mind*, vol. 97 (388): 551–574. [27], [33]

—. 2004. "Causal Models, Token Causation, and Processes." *Philosophy of Science*, vol. 71 (5): 820–832. DOI 10.1086/425057. [61], [70]

—. 2007. "Causation in Context." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Huw Price & Richard Corry, editors, chap. 8, 191–223. Clarendon Press, Oxford. [61], [103], [121]

—. 2008. "Counterfactual Theories of Causation (The Stanford Encyclopedia of Philosophy)." URL `http://plato.stanford.edu/entries/causation-counterfactual/`. [62], [70]

Ney, Alyssa. 2009. "Physical Causation and Difference-Making." *The British Journal for the Philosophy of Science*, vol. 60: 737–764. [50]

Norton, John D. 2007. "Causation as Folk Science." In *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Huw Price & Richard Corry, editors, chap. 2, 11–44. Clarendon Press, Oxford. [9]

Paul, L. A. 2000. "Aspect Causation." *Journal of Philosophy*, vol. 97 (4): 235–256. [ix], [144]

—. 2004. "Aspect Causation." In Collins et al. (2004). [ix], [144]

Paul, L. A. & Ned Hall. 2013. *Causation: A User's Guide*. Oxford University Press, Oxford. [ix], [52], [71], [144], [152], [153], [160], [169]

Pearl, Judea. 1999. "Simpson's Paradox: An Anatomy." Available at `http://bayes.cs.ucla.edu/R264.pdf`. [61]

—. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge. [61], [65]

—. 2009. "Causal Inference in Statistics: An Overview." *Statistics Surveys*, vol. 3: 96–146. [61], [67]

Price, Huw. 1992. "Agency and Causal Asymmetry." *Mind*, vol. 101 (403): 501–520. [33], [35], [36], [37]

Price, Huw & Richard Corry, editors. 2007. *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Clarendon Press, Oxford. [46]

Psillos, Stathis. 2009. "Regularity Theories." In Beebee et al. (2009), chap. 7, 131–157. [2], [3], [4]

Quine, W. V. 1986. *The Philosophy of Logic*. Harvard University Press, Cambridge, MA, second edn. [vii]

Reichenbach, Hans. 1956. *The Direction of Time*. Dover Publications, Mineola. [1]

Russell, Bertrand. 1912. "On the Notion of Cause." *Proceedings of the Aristotelians Society*, vol. 13: 1–26. [1], [9], [37], [46]

—. 1948. *Human Knowledge: Its Scope and its Limits*. Allen & Unwin, London. [18]

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton. [1], [11]

—. 1994. "Causality without Counterfactuals." *Philosophy of Science*, vol. 61 (2): 297–312. [1], [49], [50]

Sartorio, Carolina. 2006. "Disjunctive Causes." *Journal of Philosophy*, vol. 103: 521–38. [162]

Schaffer, Jonathan. 2001. "Causation, Influence, and Effluence." *Analysis*, vol. 61 (1): 11–19. [54]

—. 2004. "Trumping Preemption." In Collins et al. (2004), chap. 2, 59–75. [166]

—. 2005. "Contrastive Causation." *The Philosophical Review*, vol. 114 (3): 297–328. [169]

—. 2012. "Causal Contextualism." In *Contrastivism in Philosophy*, Blaauw, editor, chap. 2, 35–63. Routledge. [169]

Shulz, Katrin. 2011. ""If you'd wiggled A, then B would've changed": Causality and counterfactual conditionals." *Synthese*, vol. 179: 239–251. [61]

Sider, Theodore. 2011. *Writing the Book of the World*. Clarendon Press, Oxford. [vii], [3]

Smith, Sheldon R. 2007. "Causation and Its Relation to 'Causal Laws'." *The British Journal for the Philosophy of Science*, vol. 58 (4): 658–688. [9]

Sosa, Ernest & Michael Tooley, editors. 1993. *Causation*. Oxford Readings in Philosophy. Oxford University Press, Oxford.

Spirtes, Peter, Clark Glymour & Richard Scheines. 2000. *Causation, Prediction, and Search*. The MIT Press, Cambridge, ma, second edn. [61]

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In *Studies in Logical Theory*, N. Rescher, editor, chap. 4, 98–112. Oxford University Press, Oxford. [viii], [ix], [63], [71], [78], [79], [80], [144]

—. 1980. "A Defense of Conditional Excluded Middle." In *Ifs*, W. L. Harper, R. Stalnaker & G. Pearce, editors, 87–104. D. Reidel, Dordrecht. [71]

Starmans, Christina & Ori Friedman. 2012. "The Folk Conception of Knowledge." *Cognition*, vol. 124 (3): 272–283. [17]

Strevens, Michael. 2003. "Against Lewis's New Theory of Causation: A Story with Three Morals." *Pacific Philosophical Quarterly*, vol. 84 (4): 398–412. [54], [169]

—. 2008. *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA. [28], [38], [46], [47], [54], [55], [56], [57], [58]

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. North-Holland Publishing Company, Amsterdam. [58]

Thomson, Judith Jarvis. 2003. "Causation: Omissions." *Philosophy and Phenomenological Research*, vol. 66 (1): 81–103. [169]

Turri, John. 2013. "A Conspicuous Art: Putting Gettier to the Test." *Philosopher's Imprint*, vol. 13 (10): 1–16. [18]

VanderWeele, Tyler J. 2009. "Criteria for the characterization of token causation." *Logic and Philosophy of Science*, vol. VII (1): 115–127. [65]

Varzi, Achille C. & Roberto Casati, editors. 1996. *Events*. Dartmouth, Aldershot. [174]

Weslake, Brad. ms. "A Partial Theory of Actual Causation." Available at http://bweslake.s3.amazonaws.com/research/papers/weslake_ac.pdf. [61]

Williamson, Timothy. 2007. *The Philosophy of Philosphy*. The Blackwell/Brown Lectures in Philosophy. Blackwell Publishing, Malden, MA. [17], [18]

Woodward, James. 1999. "Causal Interpretation in Systems of Equations." *Synthese*, vol. 121: 199–247. [61]

—. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. [7], [50], [58], [61], [62], [67], [103]

Woodward, James & Christopher Hitchcock. 2003a. "Explanatory Generalizations, Part I: A Counterfactual Account." *Noûs*, vol. 37 (1): 1–24. [61], [62], [70]

—. 2003b. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs*, vol. 37 (2): 181–199. [61]