# Seminar Notes for *Causality*

J. Dmitri Gallow

Spring, 2017

# Contents

# 1 Some Preliminary Distinctions and Comments on Methodology

## 1.1 The Subject Matter of the Course

1. In this course, we will be interested in claims like the following:

   (a) Nixon caused the U.S. economy to stagflate in the early 1970's.

   (b) The Supreme Court's decision in *Roe v. Wade* caused crime to decrease in the U.S. in the 1990's.

   (c) Drowsy driving causes crashes.

   (d) Drinking a gallon of antifreeze causes death.

   (e) Helmet laws prevent death by preventing serious brain damage in accidents.

   (f) Helmet laws cause death by reducing visibility, thereby causing accidents.

2. Note: by focusing on these kinds of causal *claims*, we do not presuppose that each of the claims above is talking about the *very same* kind of causal relation.

   (a) It is commonly thought, e.g., that (1a) and (1b) are talking about a very different kind of causal relation than (1c) and (1d). Whereas (1a) and (1b) are describing *singular, token*, or *actual* causal relations, (1c) and (1d) are describing *general*, *type*, or *property-level* causal relations.

   (b) If we assume that "cause" and its cognates (e.g., "prevents") are univocal, then (1e) and (1f) would appear to be contradictory. However, we may wish to distinguish between *component* effects and *net* effects, and thereby allow (1e) and (1f) to both be true at once. (see Hitchcock (2001a))

   (c) We do not even presuppose that *there is* any such thing as the causal relation. It could be that, while (1a–1f) are all assertable, none of them are true. (see Strevens (2008))

3. While (1b) mentions an *event*—something that took place at a particular place and time[1]—(1a) mentions a *person*—Nixon.

[1] There's debate about whether the subject in (1b) is an *event* or a *fact* (see, e.g., Bennett (1988, 1996) and Mellor (1995)). For present purposes, let's use 'event' in a sense that it neutral between facts and events. Right now, we just want to distinguish between object-causation and more traditional kinds of causation.

(a) In the 'cause' place, English allows objects, people, and events.

   i. Suzy's throwing the ball caused the window to shatter. [EVENT]

   ii. Suzy caused the window to shatter. [PERSON]

   iii. The ball caused the window to shatter. [OBJECT]

(b) The standard line on this is that (**??**), (3(a)ii), and (3(a)iii) are, if true, true because of the very same causal relation. When we say that *Suzy* caused the window to shatter, we mean that (or: what we mean is true in virtue of the fact that) Suzy's *throwing the ball* caused it to shatter. That is: fundamentally, causation is a relation between events; statements of object causation are true in virtue of causal relations between events

(c) Interestingly (?), while English allows us to replace the event with the agent of the event in the *cause*, it does not allow us to do this with the *effect*.

   i. Suzy caused the window to shatter.

   ii. # Suzy's throwing the ball caused the window.

### 1.2 Philosophical Accounts of Causation

1. *Reductionists* about causation deny that whether $c$ caused $e$ is a *brute* fact about the world. They think that causal facts can be *reduced* to other kinds of facts—*e.g.*, facts about the Humean mosaic and the laws of nature.

2. *Anti-reductionists* about causation claim that whether $c$ caused $e$ is a brute fact about the world, not to be reduced to any other facts.

   (a) Armstrong (1997) uses cases of probabilistic causation to attempt to establish anti-reductionism. Suppose that the probability that $c_1$ cause $e$, given that $c_1$ occurs, is 0.5; and that the probability that $c_2$ cause $e$, given that $c_2$ occurs, is 0.5; and whether $c_1$ causes $e$ is independent of whether $c_2$ causes $e$. If $c_1, c_2$, and $e$ all occur, then this is compatible with $c_1$ causing $e$ and $c_2$ not, $c_2$ causing $e$ and $c_1$ not, and $e$ being overdetermined by both $c_1$ and $c_2$. So the non-causal facts do not suffice to determine the causal facts. So reductionism is false.

3. For the most part, the authors we are going to read are reductionists who are interested in providing a philosophical *theory* of causation; or a philosophical *account* of causation.

   A philosophical account of causation

   A philosophical account of causation is a claim of the following form:

   Necessarily, for all $c$ and $e$, $c$ caused $e$ iff $R(c, e)$.

   (a) Here, '$R$' is just any (binary) relation.

      i. Note: I've assumed here that causation is a binary relation. This assumption may be false; we will encounter authors later in the semester who think that causation is a three or four-place relation.

ii. Note also: we saw above that there may be many different varieties of causation—type, token, component, and net, for instance. An account like the one above could be given for one particular species of causation, and fall silent on the other species of causation.

(b) One true philosophical account of causation, then, is this one:

Necessarily, for all $c$ and $e$, $c$ caused $e$ iff $c$ caused $e$.

(c) This account is true, but it is not particularly interesting or informative. (Though, for the anti-reductionist, this is the *only* true philosophical theory of causation.)

(d) Interesting or informative accounts subdivide into two kinds: *reductive* and *non-reductive*.

A REDUCTIVE ACCOUNT OF CAUSATION
A reductive account of causation is a claim of the form:

Necessarily, for all $c$ and $e$, $c$ caused $e$ iff $R(c, e)$.

where $R$ does not itself involve any causal notions.

(e) A *reductive* account of causation is one that allows us to *reduce* causal notions to non-causal notions without remainder.

A NON-REDUCTIVE ACCOUNT OF CAUSATION
A non-reductive account of causation is a claim of the form:

Necessarily, for all $c$ and $e$, $c$ caused $e$ iff $R(c, e)$.

where the relation $R$ is picked out using causal notions.

(f) A *non-reductive* account of causation does not allow us to reduce causal notions to non-causal notions without remainder. Nevertheless, a non-reductive account may still tell us something rather interesting about causation.

i. For instance, consider: $c$ causes $e$ iff $\Pr(e \mid c \wedge k) \neq \Pr(e \mid \neg c \wedge k)$, where $k$ is every cause of $e$ besides $c$.

ii. This is non-reductive but non-trivial. If true, it tells us something interesting about the relationship between causation and chance, even though it doesn't allow us to *reduce* causation to chance.

### 1.3 Evaluating Philosophical Accounts of Causation

1. How do we evaluate a philosophical account of causation?

(a) We think about whether it says anything false.

i. Of course, it may not always be clear whether something that the account says is false; there may be disagreement over whether what the account says is false. In that case, debate about whether the account is satisfactory will focus on this subsidiary question.

(b) For instance, suppose that I suggest the following theory of causation:

Necessarily, for all $c$ and $e$, $c$ caused $e$ iff, had $c$ not occurred, $e$ would not have occurred.

(c) This account says that, in any case where there is a reliable backup, there is no causation. So, the electric company's supplying power doesn't cause the lights to be on, since there is a backup generator (and, were the electric company to cut the power, the lights would still be on). But that's false. The electric company's supplying power *does* cause the lights to be one. So the account is false.

### 1.4 Dowe on Empirical and Conceptual Analyses

1. Dowe (2000) distinguishes two tasks for philosophers of causation:

   (a) "to elucidate our normal concept of causation" [conceptual analysis]

   (b) "to discover what causation is in the objective world" [empirical analysis]

2. *Cf.* Psillos (2009):

   (a) "a theory of the meaning of causal statements"

   (b) "a theory of what causation is in the world [or] a theory about the worldly constituents of causation"

3. In particular, Dowe denies that *conceptual analysis* (the procedure outlined in the previous section) tells us anything about what causation is *in the world*.

4. Contentious claim: the distinction between the *concept* of causation and 'what causation is in the world' is largely a side show.

   (a) There is a distinction between these two, of course. However, in the relevant sense of 'concept', there is also an intimate connection.

   (b) Surely

      i. ⟨ The cat is on the mat ⟩ is true

      ii. The cat is on the mat

      are different claims; however, they are still intimately connected. (4(b)i) is true iff (4(b)ii) is.

   (c) Similarly,

      i. The concept CAUSATION applies to $c$ and $e$

      ii. The relation ⟨CAUSATION⟩ relates $c$ and $e$.

      are different claims; yet (4(c)i) is true iff (4(c)ii) is, too.

5. A rejoinder: in some good sense of 'concept', (4(c)i) holds iff (4(c)ii) does. However, so understood, facts about the concept CAUSATION cannot be discovered from the armchair. That is: our intuitive judgments in particular cases are no guide to the concept of causation.

   (a) In some good sense of 'concept', (5(a)i) is true iff (5(a)ii) is.

      i. The concept WATER applies to $x$

      ii. $x$ has the property ⟨WATER⟩.

Nevertheless, in this sense of 'concept', you cannot discover whether the concept WATER applies to $x$ through a priori reflection. Your intuitive judgments about whether $H_2O$ is water do not constitute data for an empirical theory of water.

(b) E.g., our intuitive 'energy' judgments tell us that children are more energetic than their parents. However, it would be wrong to infer from this and the fact that energy is proportional to mass than children have more mass than their parents.

(c) Similarly, the rejoinder concludes, it would be wrong to use our intuitive causation judgments to decide whether $c$ caused $e$.

6. An aside on Dowe's comparison with 'energy' (reproduced below):

> We can say that application of the scientific method of theorizing and experimentation produced an 'empirical analysis' of energy. In the same way, science may reasonably be expected to throw light on the language-independent entity called 'causation'. (p. 7)

(a) Surely the word 'energy' does not have the same meaning in (6(a)i) and (6(a)ii):

i. Energy is proportional to mass.

ii. Children have more energy than their parents.

Both of these claims are true, yet they do not entail that children have more mass than their parents. So the meaning of 'energy' must differ in (6(a)i) and (6(a)ii).

(b) Surely the same would be true of any empirical analysis of 'causation'. If it differed substantially from our pre-theoretic judgments of causation, it would be at best polysemous with 'causation'.

i. But then, the empirical analysis would not tell us anything about the language-independent entity called 'causation' *by us*. It would only tell us something about the language-independent entity called 'causation' by users of the future polysemous term.

7. The rejoinder to the rejoinder (point 5 above):

(a) 'Energy' is a theoretical term/concept. It refers, if at all, to the thing which, at the actual world, plays (near enough) the theoretical role which energy plays in our scientific theories.

i. That is to say: we use the term 'energy' with *deference* to the actual world. The intension of 'energy' depends upon which world is actual.

(b) Similarly, 'water' is a natural kind term/concept. What it refers to depends upon which natural kind is appropriately related to our tokenings of the term 'water'.

i. That is to say: we use the term 'water' with *deference* to the actual world. The intension of 'water' depends upon which world is actual.

(c) My contention: we do not use the term 'causation' with deference to the actual world. Whether $c$ caused $e$ at some possible world $w$ depends upon on what the laws of nature are at $w$ and what actually occurs at $w$. It does not additionally depend upon which world is actual.

i. In support of this claim: when we consider fully-described possible worlds (including a description of their laws), our causal judgments about what caused what at these possible worlds are often incredibly clear and seemingly not defeasible by empirical discovery. Discovering new empirical facts about our own world does not appear to shake our judgments about what caused what at these other possible worlds.

A. Of course, I am committed to the claim that this is true of *all* possible causal relations in all possible worlds. I cannot survey all the possible causal judgments to conclusively establish this claim. If you disagree, you should try to describe a causal claim in some possible world, and some way the *actual* world could be, such that the truth of the causal claim at that possible world depends upon whether or not the actual world is that way.

(d) For this reason, empirical work is not necessary to uncover the intension of CAUSATION in the same way as it is necessary to uncover the intensions of ENERGY and WATER.

8. Finally, a *tu quoque* argument against DOWE: he utilizes precisely the method of 'conceptual analysis' that he disavows.

(a) He criticizes FAIR (1979)'s theory of causation as the transfer of energy/momentum on the grounds that a spaceship moving inertially with a constant velocity does not transfer any energy/momentum. But, Dowe says, its earlier velocity is a cause of its later velocity. Where does this causal claim come from, if not DOWE's own judgment? As DOWE explicitly says, 'cause' is not a theoretical term in any physical theory, so the judgment is not the consequence of a well-confirmed scientific theory.

(b) He claims that QM 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. (p. 8, fn. 9)

Again, where does this causal claim come from, if not DOWE's own judgment?

9. This *tu quoque* argument points towards a more important point: it's difficult to know how a philosopher could theorize about causation if they *didn't* take any causal claims for granted or rely upon causal judgments in any way.

(a) As DOWE acknowledges, 'cause' is not a theoretical term in any scientific theory. So we can't rely upon any scientific theory to tell us which causal claims are true.

(b) But then, if we have no causal claims from science, and we don't trust obvious causal truths like "dropping the glass caused it to shatter", then there will be no means of *evaluating* the verdicts of our philosophical theory of causation.

   i. Suppose that you disagree with Dowe and myself, and you think that 'cause' is a theoretical term in a scientific theory.

   ii. If so, I am inclined to rehearse the argument from point (6) above, and contend that the theoretical term 'cause' is at best polysemous with our ordinary verb 'to cause'.

   iii. However, if this doesn't move you, then you should think that philosophers simply shouldn't be in the game of theorizing about causation in the first place. You should think that theorizing about causation is a task best left to the relevant scientists.

10. What's right in what Dowe says:

   (a) Work on causation should not be blind to empirical work.

   (b) There may be some interesting physical relation—call it *biff*—which relates $c$ and $e$ at the actual world when and only when $c$ causes $e$. If so, that would be interesting to know. However, the way to find that out is to first figure out which things are causally related (by constructing a philosophical theory of causation, in the manner of the previous section), and then checking to see whether, at the actual world, any such biff exists.

# 2 Regularity Theories

## 2.1 Hume

1. Perhaps the first regularity theorist was David Hume, who defined causation twice over:

   (a) "We may define a CAUSE to be 'An object precedent and contiguouus to another, and where all the objects resembling the former are plac'd in like relations of precendency to those objects, that resemble the latter.'" (HUME, THN, 170)

   (b) "A CAUSE is an object precendent and contiguous to another, and so united with it, that the idea of the one determines the mind to form the idea of the other, and the impression of the one to form a more lively idea of the other." (HUME, THN, 170)

2. Regularity theorists following Hume have picked up on (1a), and not worried about the more psychological definition given in (1b).

3. Hume speaks of *objects* as causes, rather than *events*. Following present orthodoxy, we can understand him as talking about the event of an object taking on or retaining a certain property. Then, we can render Hume's view as follows

   HUME'S THEORY (TYPE CAUSATION)
   Event type $C$ causes event type $E$ iff the occurrence of events of type $C$ is *constantly conjoined* with the occurrence of events of type $E$.

   HUME'S THEORY (TOKEN CAUSATION)
   A token event $c$ caused a token event $e$ iff there are event types $C$, $E$ such that:
   (a) $c$ is of type $C$;
   (b) $e$ is of type $E$;
   (c) event type $C$ causes event type $E$.

   (a) Note: in these definitions (and throughout), we assume that both $c$ and $e$ actually occurred.

   (b) We'll understand the phrase "constant conjunction" in such a way that:

i. it builds in temporal precedence—if events of type $C$ are constantly conjoined with events of type $E$, then events of type $C$ are constantly *prior to*, and contiguous to, events of type $E$; and

ii. it merely requires that the occurrence of events of type $C$ is *sufficient* for the occurrence of events of type $E$; it does not require that they be necessary—events of type $E$ could fail to be preceded by events of type $C$.

4. For Hume, *type* causation is more fundamental than *token* causation.

### 2.2 Mill

5. J. S. Mill:

> "This invariable sequence seldom if ever holds between a consequent and a single antecedent. It's usually between a consequent and the sum of several antecedents, the concurrence of all of them being needed to produce—*i.e.* to be certain of being followed by—the consequent. People often single out one of the antecedents as the 'cause' and call the others merely 'conditions'.
>
> ...Philosophically speaking, then, the cause is the sum total of the conditions, positive and negative—the whole of the contingencies of every sort from which the consequent invariably follows." (Mill, 1843, Book III, ch. 5, §3)

   (a) Fire isn't *always* followed by smoke; smoking isn't *always* followed by cancer.

   (b) Mill's solution is to say that, while fire is not—philosophically speaking—a cause of smoke; and while smoking is not—philosophically speaking—a cause of cancer, they are nevertheless *parts* of the *total cause* of smoke and cancer.

   (c) Chris contracts cancer. The *total cause* of this cancer will include Chris's smoking, Chris's diet, his physical constitution, the lack of sufficient medical techniques to prevent the cancer, and so on and so forth. We might *say* that the smoking caused the cancer, but, on Mill's view, this is simply incorrect.

6. Mill thus advocated (as a very rough exegesis)[1] the following theory of causation:

> Mill's Theory (Type Causation)
> The set of event types $T = \{C_1, C_2, \ldots, C_N\}$ is a *total cause* of event type $E$ iff the co-occurrence of the types of events in $T$ is *constantly conjoined* with the occurrence of events of type $E$.
>
> Mill's Theory (Token Causation)
> A collection of token events $t$ is the *total cause* of a token event $e$ iff:
>
> (a) Each event in $t$ is of a type $C_i \in T$;

[1] I've included Mill in this history primarily for expository purposes, and I am far from confident that I have all the particulars of his view correct. Take with a grain of salt.

(b) For every event type $C_i \in T$, there is some event in $t$ which is of that type;

(c) $e$ is of type $E$;

(d) The set of event types $T$ is a *total cause* of the event type $E$.

(a) We are assuming that the each event in the set of events $t$ actually occurred.

(b) Mill accepts that there may be multiple total causes of any event type.

(c) Though I've used the phrase 'total cause' to emphasize what's going on, for Mill, the total cause *is* the cause.

## 2.3 Mackie

7. For Mill, the following claim was (philosophically speaking) false:

   (a) Striking the match caused it to light.

   For striking the match was not—*by itself*—sufficient for the match to light. In addition, there must have been oxygen present; the match must have been dry, it must not have been too windy, and so on and so forth. For Mill, philosophically speaking, (7a) is on a par with (7b).

   (b) The presence of oxygen caused the match to light.

8. Mackie (1965) revises this aspect of Mill's theory. He wants to allow that (7a) is—even strictly speaking—true. He also wants to do something to explain why (7a) seems more appropriate to utter than (7b).

   (a) The guiding idea is this: a cause isn't Mill's *total cause*. Rather, it is one of the *parts* of a total cause.

      i. Note that, if *one* set of event types $T$ is a total cause of the event type $E$, then so too will be any *superset* of $T$.

      ii. Thus, if we accept Mackie's proposal as it is, then we would say that my vote for Clinton caused Trump to win. For, if there is *some* total cause $T$ of Trump's victory, then adding my vote for Clinton to that set will *still* be a total cause, according to Mill's theory. But then my vote for Clinton will be a *part* of a Millian total cause.

      iii. Instead, Mackie requires that the total cause $T$ be *minimal*—in the sense that, not only is $T$ sufficient for $E$, but *no subset of $T$ is sufficient for $E$*.

   (b) Mackie's own statement of his view often makes use of the idea of an INUS condition—an *insufficient* but *non-redundant* part of an *unnecessary* but *sufficient* condition. So Mackie first suggests this:

   Mackie's Theory (Type Causation)—First Pass
   The event type $C$ is a cause of the event type $E$ iff:

(a) There is a set of event types $T$ which is *sufficient* but *not necessary* for $E$;
(b) $T$ is minimal—no subset of $T$ is sufficient for $E$; and
(c) $C$ is included in $T$.

Mackie's Theory (Token Causation)—First Pass
The token event $c$ caused the token event $e$ iff there are types $C$, $E$, and a set of event types $T$ such that:
(a) $c$ is of type $C$;
(b) $e$ is of type $E$;
(c) $T$ is *sufficient*, but *not necessary*, for the event type $E$;
(d) $T$ is minimal—no subset of $T$ is sufficient for $E$;
(e) $C$ is included in $T$;
(f) a token of each event in $T$ actually occurred; and
(g) for no other sufficient condition for $E$, $T'$, did a token of each event type in $T'$ occur.

(c) Note: in cases of overdetermination, this account rules that no part of either sufficient condition is a cause.
   i. Billy and Suzy both throw their rocks at the window, and both hit at the same time. On this view, both of the following claims are false:
      A. Billy's throw caused the window to shatter.
      B. Suzy's throw caused the window to shatter.
   ii. However, Mackie claims that the *disjunction* of Billy's and Suzy's throw caused the window to shatter. (I don't see, however, how this follows from his account.)

9. Mackie sees reason to object to this characterization involving INUS conditions.
   (a) In the first place, it could be that the event type $C$ is—*all by itself*—sufficient for an event of type $E$. If that were so, then $C$ would not be *insufficient* on its own. However, Mackie in this case still wants to call $c$ a token cause of $e$.
   (b) In the second case, it could be that the total cause $T$ which has $C$ as a part is the *only* minimally sufficient condition for $E$. In that case, it would not be an *unnecessary* sufficient condition. However, Mackie in this case still wants to call $c$ a token cause of $e$.
   (c) In the third place, Mackie thinks there are cases of overdetermination—cases in which two minimally sufficient conditions for $E$, $T$ and $T'$, both obtain—in which the event type $C$ is in both $T$ and $T'$. In this case, too, Mackie wishes to say that $c$ is a token cause of $e$.

10. Mackie himself ends up complicating the notion of an INUS condition—replacing it with the notion of being *at least an INUS condition*. But I think it's cleaner to just present his view in the following way:

Mackie's Theory (Type Causation)
The event type $C$ is a cause of the event type $E$ iff:

(a) There is a set of event types $T$ which is sufficient for $E$;

(b) $T$ is minimal—no subset of $T$ is sufficient for $E$; and

(c) $C$ is included in $T$

Mackie's Theory (Token Causation)
The token event $c$ caused the token event $e$ iff there are types $C$, $E$, and a set of types $T$ such that:

(a) $c$ is of type $C$;

(b) $e$ is of type $E$;

(c) $T$ is sufficient for $E$;

(d) $T$ is minimal—so subset of $T$ is sufficient for $E$;

(e) $C$ is included in $T$;

(f) a token of each event type in $T$ actually occurred; and

(g) for no other sufficient condition for $E$ *not including* $C$, $T'$, did a token of each event type in $T'$ occur.

(a) To summarize this all in a slogan: $C$ caused $E$ iff $C$ is a **P**art of a **M**inimally **S**ufficient condition for $E$ (a type cause is a PMS condition for the effect). And $c$ is a token cause of $e$ iff $c$ is a **P**art of **A**ll **O**ccurrent **M**inimally **S**ufficient conditions for the effect (a token cause is a PAOMS condition for the effect).

## 2.4 Mackie on Selection and Context-Sensitivity

11. Within Mill and Mackie's framework, the so-called *problem of selection* is the problem of saying something about which of the multitudinous *parts* of total causes we *select* to cite as causes.

    (a) It is common-sensical to say that the striking of the match was a *cause* of the match's lighting, whereas the presence of oxygen was merely a *background condition* for the match's lighting.

    (b) What makes the difference between these two? On Mill and Mackie's view, the difference is *not* metaphysical, but rather something having to do with our causal *talk*.

12. These "principles of invidious discrimination" (in Lewis (1973a)'s words) appear to be context-sensitive. For an example from Hart & Honoré (1985): imagine that the Indian government does not have food reserves on hand, and there is a flood which destroys much of that year's crops. There is subsequently a famine. Then, some may say:

    (a) The flood caused the famine.

    While others may say:

(b) There was bound to be a disaster eventually. It wasn't the famine, but the Government's poor planning, which caused the flood.

The person saying (12a) is treating the government's lack of reserves as a background condition; while the person saying (12b) is treating the flood as a background condition. At first glance, it doesn't appear that (12a) (in its context) is inconsistent with (12b) (in its context). So it looks, at first glance, like principles of selection are going to be context-sensitive.

13. Note also that emphasis can make a difference to the appropriateness of causal claims. It sounds okay to say:

    (a) Socrates' drinking *Hemlock* at dusk caused him to die.

    But it sounds quite odd—and, to my ear, *false*—to say:

    (b) Socrates' drinking Hemlock *at dusk* caused him to die.

14. Mackie provides us with a theory which can explain what's going on in these cases.

    (a) When we make and evaluate causal claims, we are often taking certain things for granted. Appropriating Mackie's terminology, let's call those things we are taking for granted the *causal field*.

    (b) For Mackie, the causal field is a set of particular cases (individuals, houses, economies, match-strikings, *etc.*). When there is a particular causal field in play, this is because we are presupposing that the case of interest falls somewhere within this set.

    (c) This set will generate a set of conditions: namely, the set of conditions which each case satisfies. Call these conditions 'presupposed' by the causal field. For instance, if each match-striking in the causal field includes the presence of oxygen, then the presence of oxygen will be presupposed by the causal field.

    (d) A set of event types $T$ is minimally sufficient for an event type $E$ *within a causal field* $\mathcal{F}$, iff, in each case in $\mathcal{F}$, the occurrence of all of the event types in $T$ is invariably followed by an event of type $E$.

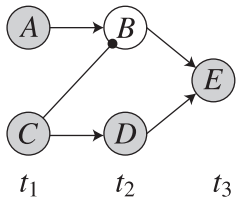
    (e) Something that is a part of a minimally sufficient condition *full stop* may not be a part of any minimally sufficient conditions *within a particular causal field*.

        i. E.g., if the causal field presupposes the presence of oxygen, then the presence of oxygen will not be a part of any minimally sufficient condition for the match lighting. If not, then it will.

    (f) Thus, by taking some feature of the world for granted—excluding those cases which lack that features the causal field—things which would have been appropriately cited as causes are no longer appropriately cited as causes. Rather, they are appropriately treated as background conditions.

        i. Thus, we may say that the person who says (12a) is not taking the presence of a disaster for granted; whereas the person who says (12b) is. These differences in context can make "The flood causes the famine" true in the context of (12a) but false in the context of (12b).

**Figure 2.1** Preemption




ii. Suppose (as I believe is not wholly implausible) that emphasizing a particular part of a sentence has the semantic function of treating the unemphasized part of the sentence as a presupposition.

   A. Then, (13a) presupposes that Socrates drank something at dusk. Within a causal field which is restricted to include only those cases in which Socrates drinks something at dusk, the drinking of Hemlock *will* be a part of a minimally sufficient condition for death. And (13a) will be true.

   B. And (13b) will presuppose that Socrates drank Hemlock. Within a causal field which is restricted to include only those cases in which Socrates drinks Hemlock, it being dusk will *not* be a part of a minimally sufficient condition for death. And (13b) will be false.

iii. So: by understanding different contexts as corresponding to different causal fields, we may give a diagnosis of the initially puzzling phenomena of sentences like (12a), (12b), (13a), and (13b).

## 2.5 Objections to Regularity Theories

15. Mackie considers some objections to his theory. However, I think that the cases he considers obscures the force of those objections. Because I think they elicit clearer intuitions and afford fewer chances for re-description, I'll make the objections using Lewis (1986a)'s neuron diagrams.

16. First objection: Preemption

    (a) Consider the neuron diagram in figure 2.1. There, *C*'s firing caused *E*'s firing. *A*'s firing did not cause *E*'s firing.

    (b) The set containing the event type { an *A*-firing } is minimally sufficient for an *E*-firing. (Any situation consistent with the neuron laws at which an *A*-firing occurs is followed shortly thereafter by an *E*-firing.)

    (c) The set containing the event type { a *C*-firing } is minimally sufficient for an *E*-firing. (Any situation consistent with the neuron laws at which a *C*-firing occurs is followed shortly thereafter by an *E*-firing.)

**Figure 2.2** A Common Cause

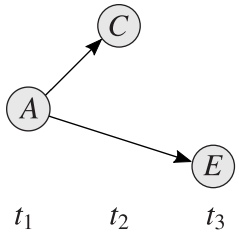
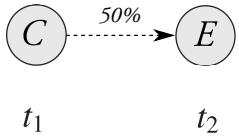



**Figure 2.3** Indeterminism




(d) Mackie's account will therefore rule that neither *A*'s firing nor *C*'s firing caused *E*'s firing.

(e) He will treat this case just like a case of *symmetric overdetermination*. But things are not symmetric in this case.

17. Second objection: Common Causes

(a) Consider the neuron diagram in figure 2.2. There, *A*'s firing caused *E*'s firing, and *C*'s firing did not cause *E*'s firing.

(b) The set containing the event type { an *A*-firing } is minimally sufficient for an *E*-firing. (Any situation consistent with the neuron laws in which an *A*-firing occurs is followed shortly thereafter by an *E*-firing.

(c) The set containing the event type { a *C*-firing } is minimally sufficient for an *E*-firing. (Any situation consistent with the neuron laws in which a *C*-firing occurs is followed shortly thereafter by an *E*-firing.

(d) Mackie's account will therefore rule that neither *A*'s firing nor *C*'s firing caused *E*'s firing.

18. Third objection: Indeterminism

(a) Consider the *indeterministic* neuron diagram shown in figure 2.3. There, the chance of *E* firing, conditional on *C* firing, is 50%. The chance of *E* firing, conditional on *C* not firing, is 0%.

(b) In this neuron diagram, *C*'s firing caused *E*'s firing (though there was some chance that it wouldn't).[2]

(c) However, *there is no* minimally sufficient condition for *E*'s firing. Half of the time when *C* fires, *E* fires, and half of the time when *C* fires, *E* doesn't fire.

(d) So Mackie's account (and, indeed, *any* account which requires sufficiency) will say that *C*'s firing didn't cause *E*'s firing.

[2] Note: this claim turns out to be contentious. Some wish to say that *C*'s firing caused *the chance* of *E*'s firing to be 50%, but that *E*'s firing itself was entirely uncaused.

# 3 What are the Causal Relata?

1. Causation is a two-place relation (we're supposing for now). What are its relata?
2. Three prominent possibilities:
   (a) Facts
   (b) Events
   (c) Variable values
      i. We'll talk more about variable values later on (they may just reduce to one of the previous two). Today, let's think about facts and events.
3. Facts are, for Mellor (1995) at least, true propositions. The identity conditions of facts will then be determined by an independent theory of propositions—for instance, we could adopt a Russellian conception of propositions, a possible-worlds conception of propositions, a hyper-intensional conception of propositions (though it's unclear that hyper-intensions are really necessary to do any work in a theory of causation), or something else altogether.
   (a) Some object to the idea that abstract entities like propositions can literally *cause* anything. Causation is thought to be a more worldly affair.
   (b) Davidson (1967) does not believe that facts are well-suited to be the causal relata because he does not believe that there are enough of them. He presents his 'slingshot' argument that there is only one fact to establish this conclusion. Take any two true propositions, $p$ and $q$. Then,

$$\begin{array}{ll} \text{P1.} & p = \langle \{x \mid x = x \wedge p\} = \{x \mid x = x\} \rangle \\ \text{P2.} & \{x \mid x = x \wedge p\} = \{x \mid x = x \wedge q\} \\ \hline \text{C1.} & p = \langle \{x \mid x = x \wedge q\} = \{x \mid x = x\} \rangle \\ \text{P3.} & q = \langle \{x \mid x = x \wedge q\} = \{x \mid x = x\} \rangle \\ \hline \text{C2.} & p = q \end{array}$$

   (Here, C1 follows from P1 and P2 by the principle of intersubstitutivity of identicals *salva veritatae*.)
4. What, then, are events?

(a) An important aspect of our theory of events will be the *fineness of grain* of events on that theory. A theory of events is more *coarse-grained* if it identifies more events picked out by different descriptions. That is, if one theory regards "Sebastian's stoll" and "Sebastian's leisurely stroll" as picking out the very same event, then it is more coarse-grained (with respect to this pair of event-descriptions, at least) than a theory which regards them as picking out two distinct events.

### 3.1 Quine and Davidson

5. On the view commonly attributed to Quine (1985), an event was individuated by the region of spacetime in which it occurs.

   (a) Consider the following example, originally from Davidson: a ball rotates and, at the same time, heats up. Both occur in the same region of spacetime, so the ball's rotating is, on this theory, identical to its heating up. There are not two events, the rotating and the heating up, but rather just one.

   (b) Quine says "I am not put off by the oddity of such identifications. Given that its heating up warms its surroundings, I concede that its rotating, in this instance, warms the surroundings. I am content likewise to conclude that Sebastian's gum-chewing got him across Bologna, if it coincided with his walk. These results seem harmless to science, for they imply no causal connection between warming and rotation in general."

6. This theory of events is incredibly *coarse-grained*.

7. Davidson, too, ended up accepting a theory of events like this.

   (a) For Davidson, it is important that we distinguish *the event* from *the description by means of which we pick out* the event.

   (b) This is, however, a trivial point. Everyone will identify "the election of Trump" and "my least favorite event of last year". What really does the work for Davidson is his underlying theory of events, which is as coarse-grained as Quine's.

   (c) For this reason, Davidson ends up countenancing odd causal claims like the ones above, as well as:

      i. The detour caused my arrival.

      After all, the detour caused my *late* arrival; and my late arrival *just was* my arrival. So, if the detour caused one of these, then it must have caused the other as well.

8. The causal claims to which the Quine-Davidson theory appears to commit us are odd, and many have wished to adopt a more fine-grained theory of events in order to avoid them. However, there is another option, adopted by Anscombe (1969): claim that causation is an intensional relation.

(a) Then, we could say that, even though the ball's spinning and the ball's warming up are identical, still, the ball's warming up caused its surroundings to warm, and its spinning did not.

### 3.2 Lewis

9. On Lewis's theory of events, they are not just spacetime regions, as Quine and Davidson thought, but rather *classes of spacetime regions at worlds*.

   (a) For an event to *occur* at a time and place in a world is just for it to have a *member* at that world which overlaps with that time and place.

10. On this theory, events are much finer-grained than they are on the Quinean account. We can distinguish the ball's rotating and its heating up by distinguishing their other-worldly members.

11. Does any class of spacetime regions at worlds get to count as an event?

    (a) Lewis: No. And a good thing, too, for given his counterfactual theory of causation and his semantics for counterfactuals, it would follow that the necessarily omnipresent event—the event containing every region of spacetime in every world—causes *everything*.

    (b) Which classes of spacetime regions at worlds do count? Lewis does not tell us.

### 3.3 Kim

12. For Kim (1976), events are individuated by a tuple of:

    (a) A property (or relation);

    (b) An object (or tuple of objects); and

    (c) A time (or interval of time)

13. Thus, the event of the ball rotating at $t$ will be the tuple

    $$< \text{is rotating, the ball}, t >$$

    and this will be distinct from the event of the ball heating up, which will be

    $$< \text{is heating up, the ball}, t >$$

14. Which properties get to be included?

    (a) Kim: not all. We must rule mere Cambridge properties out.

i. If we allowed in mere Cambridge properties, then (depending upon the particulars of our account of causation) we could end up with my now being such that the Treaty of Versailles was signed over 100 years ago—that is,

< is such that 100 years ago the Treaty of Versailles was signed, Dmitri, 2017 >

causing the 2nd World War. This would be a bad case of backwards causation.

(b) Which ones get included, then? Kim does not tell us.

15. A common objection to both Kim and Lewis's theory is that it bloats our ontology by acknowledging many more events than is commonsensical.

(a) There is not just one stabbing of Caesar by Brutus. There is the gentle stabbing, the stabbing with a knife, the stabbing with a knife on Tuesday, and so on and so forth.

# 4 Probabilistic Theories

## 4.1 Probability and Its Interpretations

### 4.1.1 The Theory of Probability

1. The mathematical theory of probability begins with the notion of a *probability space*. A probability space contains three things:

   (a) A set of possibilities, $\mathscr{W}$;

   (b) A set of *propositions*/*events* $\mathscr{A}$ (subsets of $\mathscr{W}$);[1] and

   (c) A probability function, Pr, which is a function from $\mathscr{A}$ to $[0, 1]$.

2. Not just any function from $\mathscr{A}$ to $[0, 1]$ counts as a probability function, It must additionally satisfy two axioms: for any disjoint $A_1, A_2, A_2, \cdots \in \mathscr{A}$,

$$\Pr(\mathscr{W}) = 1 \tag{A1}$$

$$\Pr\left(\biguplus_i A_i\right) = \sum_i \Pr(A_i) \tag{A2}$$

   (a) Note: there's dispute about (A2) (known as 'countable additivity'), as it is incompatible with the existence of a fair infinite lottery. Some authors, therefore, in some applications, prefer to get by with the strictly weaker (A3) (known as 'finite additivity'): for any disjoint $A, B$,

$$\Pr(A \uplus B) = \Pr(A) + \Pr(B) \tag{A3}$$

3. Consequences of these axioms:

   (a) For any $A$, $\Pr(\overline{A}) = 1 - \Pr(A)$

   (b) For any $A, B$, $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(AB)$

   (c) For any $A$ and any partition $\{B_i\}$ of $\mathscr{W}$, $\Pr(A) = \sum_i \Pr(AB_i)$

   (d) Any probability function is representable as a *muddy Venn diagram*.

4. As a notational matter, we also introduce the following two *definitions*.

[1] It is standardly assumed that $\mathscr{A}$ be a *sigma-algebra*. $\mathscr{A}$ is a sigma algebra iff it is closed under complementation and countable union. That is: if $A \in \mathscr{A}$, then $\overline{A} \in \mathscr{A}$; and, if $A_1, A_2, \cdots \in \mathscr{A}$, then $\bigcup_i A_i \in \mathscr{A}$.

(a) We say that $A$ and $B$ are *probabilistically independent*, according to Pr, iff

$$\Pr(AB) = \Pr(A) \cdot \Pr(B)$$

(b) The probability of $A$, *given that* $B$, is

$$\Pr(A \mid B) \stackrel{\text{def}}{=} \frac{\Pr(AB)}{\Pr(B)}$$

5. Consequences of these definitions (together with the axioms):

(a) For any $A, B$, $\Pr(AB) = \Pr(A) \cdot \Pr(B \mid A)$ (if '$\Pr(B \mid A)$' is defined).

(b) $A$ and $B$ are probabilistically independent iff:

$$\Pr(A \mid B) = \Pr(A) \quad \text{and} \quad \Pr(B \mid A) = \Pr(B) \quad \textit{(if defined)}$$

(c) Bayes' Theorem:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)}{\Pr(B)} \cdot \Pr(A)$$

(d) The theorem of total probability: for any partition $\{B_i\}$ of $\mathscr{W}$,

$$\Pr(A) = \sum_i \Pr(A \mid B_i) \cdot \Pr(B_i)$$

### 4.1.2 Interpretations of Probability Theory

1. Many different structures satisfy these axioms.

(a) Let $\mathscr{W}$ be the unit square, let $\mathscr{A}$ be subsets of that square, and let Pr be *area*.

(b) Let $\mathscr{W}$ be any finite set, let $\mathscr{A}$ be $\wp(\mathscr{W})$, and let $\Pr(A)$ be $\#A/\#\mathscr{W}$.

2. However, most of these structures are not what we are intuitively thinking of when we talk about *probability*. Even assuming that the axioms (A1) and (A2) (or, at least, (A1) and (A3)) are true of the thing(s) we are intuitively thinking of when we talk about probability, there is still a philosophical project of explicating what *else* we mean when we talk about probability—what it takes for a probabilistic claim to be *true*.

(a) That is: we are interested in a theory of the following sort:

A Philosophical Interpretation of Probability

the probability of $A$ is $x$ iff _____

(b) The probability axioms will constitute a *constraint* on such a theory—that is, the theory had better end up being one on which the axioms are true.

(c) Just as with the notion of causation, there may be *multiple* different senses of probability; therefore, different interpretations of probability need not be competitors.

3. It is generally accepted nowadays (at least by philosophers) that there are at least two importantly different kinds of probability claims: *subjective* and *objective*.

   (a) To clearly distinguish between the two, philosophers generally use 'credence' to refer to the subjective probability, and use 'chance' to refer to objective probability.

   (b) When it comes to probabilistic theories of causation, people are not generally thinking about the probabilities as credence. So we'll be focusing our attention on interpretations of chance.

## Frequentism

1. A very common attempt at explicating the notion of chance:

   ACTUAL FREQUENTISM
   The chance of an outcome $O$, in a trial of kind $K$, is $x$ iff the actual frequency of outcome $O$, in $K$-trials, is $x$.

   $$\frac{\text{the total \# of actual outcomes of kind } O \text{ in } K \text{ trials}}{\text{the total \# of actual } K \text{ trails}} = x$$

   (a) Note: everyone will accept that actual frequencies provide *evidence* about the chances. The finite frequentist goes further by *identifying* chances with actual frequencies.

   (b) Note: this is an account of *type* chance—the chance of getting this *type* of outcome, given this *type* of experimental set up. It is not an account of *token* chance—the chance that this *token* experimental set up have any particular outcome.

2. A big problem for actual frequentism is the so-called *problem of the single case*. If only one $K$-trial ever occurs, then it cannot have a non-trivial chance of coming out other than it actually did. But it seems that, *e.g.*, a fair coin can be flipped only once.

3. Another historically prominent for actual frequentism is the so-called *reference class problem*.

   (a) One of the largest roles that chances play in our cognitive lives is in helping us form expectations about, and plans for, the future. For instance, expectations about, and plans for, my future health.

   (b) In order to form these expectations and plans, I will want to know something about the *token* chance that *I* will contract cancer.

   (c) However, due to the fact that actual frequentism determines type chances by looking at actual frequencies within a certain *reference class*, and there are several such reference classes into which I fall, there will be multiple different type chances that I will contract cancer.

4. One final problem is what we can call an *explanatory* problem for actual frequentism.

(a) If one wants to *explain* why the actual frequency of heads landings is about 1/2, one should appeal to the fact that the *chance* that a flipped coin land heads is about 1/2.

(b) But, according to the actual frequentist, the fact that the chance that a flipped coin lands heads is about 1/2 *just is* the fact that the actual frequency of heads landings is about 1/2.

5. Another quite prominent form of frequentism is *hypothetical* frequentism.

> HYPOTHETICAL FREQUENTISM
> The chance of an outcome $O$, in a trial of kind $K$, is $x$ iff the frequency of outcome $O$ would approach $x$ in the limit as the number of $K$-trials approached $\infty$.
>
> $$\Pr_K(O) = x \iff \lim_{n\to\infty} \frac{\#\text{ of outcomes } O \text{ in first } n\ K\text{-trials}}{n} = x$$

(a) Note: this is *not* a consequence of the probability axioms. What *is* a consequence of the probability axioms (the so-called 'strong law of large numbers') is that, *if* the probability of $O$ in a $K$-trial is $x$, *then*,

$$\Pr\left(\lim_{n\to\infty} \frac{\#\text{ of outcomes } O \text{ in first } n\ K\text{-trials}}{n} = x\right) = 1$$

6. This account, if successful, solves the problem of the single case. It does not solve the reference class problem or the explanatory problem.

7. Nevertheless, it faces new objections which the actual frequentist account does not face. Here's just one of those new objections:

(a) If the chance of a flipped coin landing heads is 1/2, then, if the coin is flipped infinitely many times, it *might* lands heads every single time.

(b) If it $C$ might have been false, were $A$ true, then it is not the case that, if $A$ were true, then $C$ would be true. (This is known as 'The Duality Thesis', and it's very controversial.)

$$(A \diamond\!\!\rightarrow \neg C) \;\Rightarrow\; \neg(A \mathrel{\square\!\!\rightarrow} C)$$

(c) So, if the chance of a flipped coin landing heads is 1/2, then it is not the case that, were the coin flipped infinitely many times, the limiting frequency of heads would be 1/2.

## Propensitism

1. Another prominent interpretation of chances has it that they are the *propensities* of a certain chance set-up to produce certain outcomes. For lack of a better term, let's call this view 'propensitism'.

2. According to propensitism, chances are dispositional properties of repeatable chance set-ups. There are two forms of propensity theory, corresponding to two different kinds of dispositions.

Long-Run Propensitism
The chance of an outcome $O$, in a particular chance set-up $K$, is $x$ iff a chance set-up $K$ has a disposition to produce the limiting frequency $x$ of $O$ outcomes when trials of kind $K$ are repeated infinity many times.

(a) This is a view of *token* chances

3. This account runs into a different explanatory problem.

   (a) One would like to explain why the long-run frequencies are what they are by citing the chances.

   (b) But, according to the propensitist, a chance just is a disposition to produce precisely those long-run frequencies.

   (c) One cannot explain a phenomenon by citing a disposition to produce that very phenomenon—*cf.* Moliere's "dormative virtues".

   (d) So, on the propensitist view, chances cannot explain long-run frequencies.

4. Another version of propensitism says that chances are dispositions, *not* to produce certain long-run frequencies, but rather dispositions, of various strengths, to produce *particular* outcomes in the single case.

   Single Case Propensitism
   The chance of an outcome $O$, in a particular chance set-up $K$, is $x$ iff the chance set-up $K$ has a disposition of strength $x$ to produce the outcome $O$.

5. An objection to single-case propensitism: why should we think that these dispositions obey the probability calculus?

## 4.2 A Naive Probabilistic Theory of Causation

1. To understand the nuances of Suppes (1970)'s and Eells (1991)'s theory, let's begin by considering a much *worse* probabilistic theory of causation—what we can call "the naive account".

   The Naive Probabilistic Account of Causation
   $C$ caused $E$ iff
   $$\Pr(E \mid C) > \Pr(E)$$

   (a) According to the naive account, causation is just probability raising.

      i. Depending upon the interpretation of probability we have in mind, we could be talking about *token* causation (if the probabilities in question are *token* chances) or *type* causation (if the probabilities in question are *type* chances).

   (b) We could instead have written '$\Pr(E \mid C) > \Pr(E \mid \overline{C})$'. For it is a theorem of the probability calculus that
   $$\Pr(E \mid C) > \Pr(E \mid \overline{C}) \iff \Pr(E \mid C) > \Pr(E)$$

(c) A terminological point: we may wish to say that $C$ caused $E$ iff $C$ raises the probability of $E$, and then say that $C$ *prevents* $E$ iff $\overline{C}$ causes $E$—which will be so, iff $C$ *lowers* the probability of $E$. Alternatively, we may wish to see *both* probability raising *and* probability lowering as species of causation, and write that $C$ caused $E$ iff $\Pr(E \mid C) \neq \Pr(E)$. We could then say that $C$ *promotes* $E$ iff $C$ raises the probability of $E$, and that $C$ *prevents* $E$ iff it lowers the probability of $E$.

i. From my perspective, this is just a conventional choice; and not much of interest hangs on the convention we choose.

### 4.2.1 Objections to the Naive Account

2. Why shouldn't we accept the naive account? It faces at least three prominent counterexamples (one of which we've seen already in the course). There are other problems with the account, but let's just focus on these three here.

(a) Objection 1: Causation is asymmetric, but probability raising is symmetric.

i. It follows from the axioms of probability and the definition of conditional probability that

$$\Pr(E \mid C) > \Pr(E) \iff \Pr(C \mid E) > \Pr(C)$$

ii. But effects do not cause their causes. So the naive account cannot be correct.

iii. The standard solution to this objection is to simply require that causes precede their effects (as Suppes does). This leaves us unable to deal with causes of backwards causation, but probabilistic theorists are willing to pay that price (at least provisionally).

(b) Objection 2: not all probability-raisers are causes (the right-to-left hand direction of the account is false).

i. Let '$B$' be the proposition that the barometer indicates a storm, and let '$S$' be the proposition that there is a storm. Then,

$$\Pr(S \mid B) > \Pr(S)$$

and the barometer reading precedes the storm. Still, the barometer reading does not *cause* the storm.

(c) Objection 3: not all causes are probability raisers (the left-to-right hand direction of the account is false).

i. First example (attributed to Deborah Rosen): you hit the golf ball off of the tree. The golf ball rebounds off of the tree and (miraculously!) you end up getting a hole-in-one. The probability that you would get a hole-in-one, given that the golf ball hit the tree, is lower than the unconditional probability that you would get a hole-in-one. Still, your hitting the golf ball off the tree did cause you to get a hole-in-one.

ii. Second example (from Hesslow (1976)): Birth control ($B$) causes thrombosis($T$). However, birth control also prevents pregnancy ($P$), which causes thrombosis. It could work out that the probability of thrombosis is unaffected by whether you have taken birth control,

$$\Pr(T \mid B) = \Pr(T)$$

even though $\Pr(T \mid PB) > \Pr(T \mid P)$, and also $\Pr(T \mid \overline{P}B) > \Pr(T \mid \overline{P})$. (This is an instance of 'Simpson's Paradox'.)

A. Note: (2(c)i) is an example of *token* causation without probability raising; whereas (2(c)ii) is an example of *type* causation without probability raising. Examples like (2(c)ii) are also examples of causation with probabilistic *independence*; whereas examples like (2(c)i) are examples of causation with probability *lowering*. So they pose different kinds of problems for a theory of probabilistic causation.

### 4.3 Suppes' Account

3. Suppes's account is less naive than the naive account. In the construction of his account, he is primarily concerned to address objections 1 and 2.

   (a) Objection 1 (the symmetry of probability raising) is dealt with in the standard way: by requiring that causes precede their effects.

   (b) Objection 2 (that not all probability raisers are causes) is dealt with in an ingenious way: he first lets in every probability raiser as a *prima facie* cause, and then filters out the ones which are *spurious*.

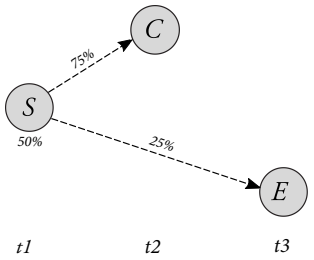
   Suppes' Probabilistic Account of Causation (v1)
   A token event $C_{t'}$ caused a token event $E_t$ iff:[2]
   (a) $t' < t$;
   (b) $\Pr(E_t \mid C_{t'}) > \Pr(E_t)$; and
   (c) there is no event $S_{t''}$ such that:
      i. $t'' < t' < t$;
      ii. $\Pr(E_t \mid C_{t'}S_{t''}) = \Pr(E_t \mid S_{t''})$
      ($S_{t''}$ screens $E_t$ off from $C_{t'}$); and
      iii. $\Pr(E_t \mid C_{t'}S_{t''}) \geqslant \Pr(E_t \mid C_{t'})$
      ($C_{t'}$ does not screen $E_t$ off from $S_{t''}$).

   (a) The terminology of "screening off" comes from Reichenbach (1956). In general, $S$ "screens off" $A$ from $B$ iff, though $A$ and $B$ are probabilistically dependent, they are probabilistically independent, conditional on $S$.

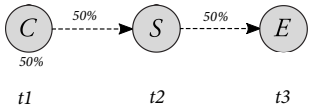
   (b) Any event meeting (a) and (b) in the account above is a *prima facie* cause of $E_t$. Any event which additionally meets (c) is a *genuine* cause. An event which meets (a) and (b) but not (c) is a *spurious cause*.

[2] Reminder: we are supposing, as usual, that both $C_{t'}$ and $E_t$ both occur.

**Figure 4.1** A screening-off common cause and a screening-off causal intermediary.




(a) A Common Cause




(b) A Causal Intermediary

4. To see how the account deals with Objection 2 to the naive account, think about the indeterministic neuron diagram shown in figure 4.1(a).

   (a) Here's how to think about these indeterministic neuron diagrams. The number below neuron $S$ in figure 4.1(a) is the probability that $S$ will fire at $t1$. The numbers above the arrows indicate the probability that the neuron at the head of the arrow will fire, *conditional on* the neuron at the base of the arrow firing; if the neuron at the base of the arrow doesn't fire, then the probability that the neuron at the head of the arrow fires is 0.

      i. We can think of each path as having a certain probability of being a *dead* or a *live* path. If the path is live, then it will act like a normal path in a neuron diagram. If the path is dead, then the neuron at the head of the path won't fire, no matter whether the neuron at the base fires or not. The numbers along each path give its probability of being a live path.

   From these numbers, we can generate the entire joint probability distribution $\Pr(\pm S, \pm C, \pm E)$ over each possible combination of neuron firings by assuming that:[3]

$$\Pr(\pm S, \pm C, \pm E) = \Pr(\pm S) \cdot \Pr(\pm C \mid \pm S) \cdot \Pr(\pm E \mid \pm S)$$

   (where '$\pm S$' is either the event of $S$'s firing or the event of its failure to fire at its designated time, and likewise for '$\pm C$' and '$\pm E$'). Then, we will generate the following distribution:

| | $\Pr(\pm S) \cdot \Pr(\pm C \mid \pm S) \cdot \Pr(\pm E \mid \pm S)$ | $\Pr(\pm S, \pm C, \pm E)$ |
|---|---|---|
| $SCE$ | $1/2 \cdot 3/4 \cdot 1/4$ | $3/32$ |
| $SC\overline{E}$ | $1/2 \cdot 3/4 \cdot 3/4$ | $9/32$ |
| $S\overline{C}E$ | $1/2 \cdot 1/4 \cdot 1/4$ | $1/32$ |
| $S\overline{C}\overline{E}$ | $1/2 \cdot 1/4 \cdot 3/4$ | $3/32$ |
| $\overline{S}CE$ | $1/2 \cdot 0 \cdot 0$ | $0$ |
| $\overline{S}C\overline{E}$ | $1/2 \cdot 0 \cdot 1$ | $0$ |
| $\overline{S}\overline{C}E$ | $1/2 \cdot 1 \cdot 0$ | $0$ |
| $\overline{S}\overline{C}\overline{E}$ | $1/2 \cdot 1 \cdot 1$ | $16/32$ |

[3] Henceforth, I'll be dropping the time subscripts for each of presentation.

(b) Then, even though $C_{t2}$ is a *prima facie* cause of $E_{t3}$, since

$$\Pr(E) = \frac{4}{32} = \frac{1}{8}$$

while

$$\Pr(E \mid C) = \frac{\Pr(C, E)}{\Pr(C)} = \frac{3/32}{12/32} = \frac{1}{4}$$

(c) $C_{t2}$ is not a *genuine* cause of $E_{t3}$—rather, it is a *spurious* cause of $E_{t3}$, since there is the event $S_{t1}$ which screens off $C_{t2}$ from $E_{t3}$,

$$\Pr(E \mid SC) = \frac{\Pr(SCE)}{\Pr(SC)} = \frac{3/32}{12/32} = \frac{1}{4}$$
$$\Pr(E \mid S\overline{C}) = \frac{\Pr(S\overline{C}E)}{\Pr(S\overline{C})} = \frac{1/32}{4/32} = \frac{1}{4}$$

and, moreover, $C_{t2}$ does *not* screen off $S_{t1}$ from $E_{t2}$, since (referencing the calculations above):

$$\Pr(E \mid SC) = \frac{1}{4} > \frac{1}{8} = \Pr(E \mid C)$$

5. Note that the temporal precedence requirement (c(i)) is crucial. Distinguish the two causal structures shown in figure 4.1. We want to say that $C$'s firing caused $E$'s firing in figure 4.1(b), but not in figure 4.1(a).

(a) However, in figure 4.1(b), $S_{t2}$ screens $E_{t3}$ off from $C_{t1}$, and $C_{t1}$ does not screen $E_{t3}$ off from $S_{t2}$.

i. In the same manner as before, figure 4.1(b) generates the joint distribution shown below.

| | $\Pr(\pm C, \pm S, \pm E)$ |
|---|---|
| $CSE$ | 1/8 |
| $CS\overline{E}$ | 1/8 |
| $C\overline{S}E$ | 0 |
| $C\overline{S}\,\overline{E}$ | 2/8 |
| $\overline{C}SE$ | 0 |
| $\overline{C}S\overline{E}$ | 0 |
| $\overline{C}\,\overline{S}E$ | 0 |
| $\overline{C}\,\overline{S}\,\overline{E}$ | 4/8 |

ii. $C_{t1}$ is a *prima facie* cause of $E_{t3}$, since

$$\Pr(E \mid C) = \frac{\Pr(C, E)}{\Pr(C)} = \frac{1/8}{4/8} = 1/4$$

and

$$\Pr(E \mid \overline{C}) = 0$$

iii. However, $S_{t2}$ *does* screen off $C_{t1}$ from $E_{t3}$, since

$$\Pr(E \mid CS) = \frac{\Pr(CSE)}{\Pr(CS)} = \frac{1/8}{2/8} = \frac{1}{2} = \Pr(E \mid S)$$

iv. And $C_{t1}$ does *not* screen $S_{t2}$ off from $E_{t3}$, since

$$\Pr(E \mid C) = \frac{1}{4} < \frac{1}{2} = \Pr(E \mid CS)$$

(b) So: Suppes *really needs* the temporal requirement that the screening-off event $S_{t''}$ *precedes* the putative cause $C_{t'}$—(c(i)).

6. Suppes then goes on to offer a second account:

> Suppes' Probabilistic Account of Causation (v2)
> A token event $C_{t'}$ caused a token event $E_t$ iff:
> (a) $t' < t$;
> (b) $\Pr(E_t \mid C_{t'}) > \Pr(E_t)$; and
> (c) there is no partition of possible events $\{K^i_{t''}\}$ such that:
> i. $t'' < t' < t$;
> ii. for each $S^i_{t''}$ in the partition,
>
> $$\Pr(E_t \mid C_{t'} K^i_{t''}) = \Pr(E_t \mid K^i_{t''})$$
>
> ($K^i_{t''}$ screens $E_t$ off from $C_{t'}$).

(a) Another, equivalent presentation of this account (one more in keeping with subsequent authors' presentations), is this: $C_{t'}$ caused $E_t$ iff:

i. $t' < t$;
ii. $\Pr(E_t \mid C_{t'}) > \Pr(E_t)$; and
iii. for every partition of possible events $\{K^i_{t''}\}$ such that $t'' < t' < t$, $C_t$ raises the probability of $E_t$ in at least one cell of this partition.

7. What is the difference between these accounts? Here's a first-pass thought which is *not* correct: the difference between the accounts comes out when there is more than one common cause of $C$ and $E$.

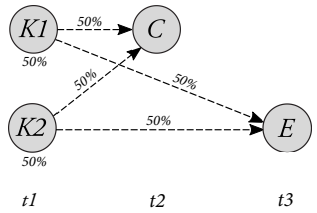
(a) For instance, consider the indeterministic neuron diagram shown in figure 4.2.[4]

(b) There, as you may verify for yourself,

$$\begin{aligned} \Pr(E \mid C) &> \Pr(E) \\ \Pr(E \mid CK1) &> \Pr(E \mid K1) \\ \Pr(E \mid CK2) &> \Pr(E \mid K2) \end{aligned}$$

[4] In this neuron diagram, suppose that the probabilities via the various paths are independent, so that $\Pr(C \mid K1K2) = 3/4$, while $\Pr(C \mid K1\overline{K2}) = \Pr(C \mid \overline{K1}K2) = 1/2$ (and likewise for $E$).

**Figure 4.2** $C_{t2}$ and $E_{t3}$ have two common causes.




(c) However, this first-pass thought is misguided. For Suppes (1970) assures us on page 38 that, if $K1$ occurs and $K2$ occurs, then there is an event $K1K2$ which occurs. And *this* prior event will screen $C$ off from $E$,

$$\Pr(E \mid C(K1K2)) = \Pr(E \mid K1K2)$$

8. The real difference between these accounts comes out, interestingly, in cases of pre-emption.

(a) Consider the neuron diagram in figure 4.3. According to Suppes's first account, $C_{t2}$ is a spurious cause of $E_{t4}$, since, even though it raised the probability of $E_{t4}$,

$$\Pr(E \mid C) = \frac{3}{4} > \frac{1}{4} = \Pr(E \mid \neg C)$$

there is an event, $K_{t1}$, which screens off $C_{t2}$ from $E_{t4}$:

$$\Pr(E \mid CK) = \frac{1}{2} = \Pr(E \mid K)$$

(b) However, according to Suppes's second account, $C_{t2}$ is a *genuine* cause of $E_{t4}$. For $C_{t2}$ both raises the probability of $E_{t4}$ *and*, conditional on one member of the partition $\{K_{t1}, \overline{K}_{t1}\}$—specifically, $\overline{K}_{t1}$—$C_{t2}$ raises the probability of $E_{t4}$:

$$\Pr(E \mid C\overline{K}) = 1 > 0 = \Pr(E \mid \overline{CK})$$

## 4.4 Objections to Suppes' Account

9. As I see things, the central difficulty with Suppes's account is that it does nothing to address Objection 3 to the naive account.

10. Just to get a more filled-out version of the objection, consider the indeterministic neuron diagram shown in figure 4.4. (This is almost precisely the same as figure 4.3, except with some minor differences in time and notation).

**Figure 4.3** A case of indeterministic preemption. $C_{t2}$ would have caused $E_{t4}$, but it was preempted by $K_{t1}$, which caused $E_{t4}$, though there was some chance that it wouldn't.

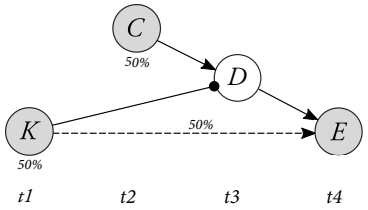
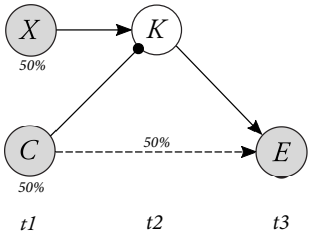



**Figure 4.4** $C_{t1}$ is probabilistically independent of $E_{t3}$, yet $C_{t1}$ caused $E_{t3}$.




(a) This indeterministic neuron diagram yields the following probability function,

| | $\Pr(\pm C, \pm K, \pm E)$ |
|---|---|
| $CKE$ | 0 |
| $CK\overline{E}$ | 0 |
| $C\overline{K}E$ | 1/4 |
| $C\overline{KE}$ | 1/4 |
| $\overline{C}KE$ | 1/4 |
| $\overline{C}K\overline{E}$ | 0 |
| $\overline{CK}E$ | 0 |
| $\overline{CKE}$ | 1/4 |

(b) Then, as you may calculate for yourself,

$$\Pr(E \mid C) = \frac{1}{2} = \Pr(E)$$

So $C_{t1}$ is not even a *prima facie* cause of $E_{t3}$; and SUPPES's account will rule, incorrectly, that $C_{t1}$ did not cause $E_{t3}$.

## 4.5 Eells's Theory

11. Ellery Eells (1991) provides us with a non-reductionist probabilistic theory of causation. On Eells's theory, type and token causation must be treated separately, and require very different accounts. Both involve the basic idea of a *probability increase*, but the interpretation of these notions and the particulars of the accounts, vary wildly from the type case to the token case.

12. Because Eells relies upon a hypothetical frequentist interpretation of probability—though he stops short of accepting such an interpretation—he inherits that theory's *reference class problem*. Therefore, for Eells, all probability claims are made relative to a *kind*[5]—this is the kind relative to which the limiting frequencies are defined. In presenting the theory, we will be holding this kind fixed; and assuming that we have a particular probability function defined already. However, the complication is relevant to Eells's handling of Hesslow (1976)'s birth-control case.

### 4.5.1 Type Causation

13. Eells's theory of type causation relies upon the idea of a *causal background context* (for $C$'s causal relevance to $E$). Though this idea is tweaked in Chapter 3, the basic idea (which we'll run with here) is this: A causal background context is defined—for a particular candidate cause $C$, and a particular effect type $E$—as follows:

    (a) Take all of the event types (or 'factors') which are causally relevant for $E$ *independently of* $C$ (excluding $C$, if $C$ is indeed causally relevant for $E$): $F_1, F_2, \ldots, F_N$.

        i. '*independently of* $C$' means that we do not include factors to which $C$ is causally relevant, nor (since Eells denies the transitivity of type causation) those factors to which the factors $C$ is causally relevant to are causally relevant, and so on and so forth.

    (b) Look at every logically possible conjunction of (negations of) these $F_i$'s: $K_1, K_2, \ldots, K_M$.

    (c) If $\Pr(CK_i) > 0$ and $\Pr(\overline{C}K_i) > 0$, then $K_i$ is a causal background context for determining $C$'s causal relevance for $E$.

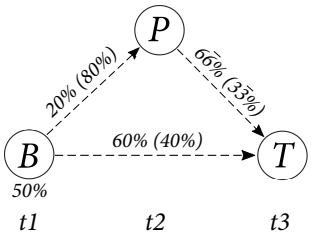
        i. For instance, suppose that there are only three (other) causal factors for $E$, excluding $C$ (if $C$ is indeed a causal factor for $E$): $F_1, F_2, F_3$. Then, there will be 8 conjunctions of (negations of) these factors:

$$\begin{aligned} K_1 &= F_1F_2F_3 & K_5 &= \overline{F}_1F_2F_3 \\ K_2 &= F_1\overline{F}_2F_3 & K_6 &= \overline{F}_1\overline{F}_2F_3 \\ K_3 &= F_1F_2\overline{F}_3 & K_7 &= \overline{F}_1F_2\overline{F}_3 \\ K_4 &= F_1\overline{F}_2\overline{F}_3 & K_8 &= \overline{F}_1\overline{F}_2\overline{F}_3 \end{aligned}$$

        Each of these which is such that both $\Pr(CK_i)$ and $\Pr(\overline{C}K_i) > 0$ is then a causal background context for determining $C$'s causal relevance for $E$

[5] A kind of population, for Eells, though we can ignore that complication here.

**Figure 4.5** An indeterministic neuron diagram. (*cf.* Hesslow (1976))




14. The theory of type causation is then the following:

    Eells's Theory of Type Causation
    An event type (or 'factor') $C$ is causally relevant to an event type $E$ iff, $C$ is *temporally prior* to $E$, and, for some *causal background context* $K_i$

    $$\Pr(E \mid C, K_i) \neq \Pr(E \mid \overline{C}, K_i)$$

    (a) If $\Pr(E \mid C, K_i) > \Pr(E \mid \overline{C}, K_i)$ for all $i$, then $C$ is a *positive* causal factor for $E$.
    (b) If $\Pr(E \mid C, K_i) \geqslant \Pr(E \mid \overline{C}, K_i)$ for all $i$, then $C$ is a *Pareto positive* causal factor for $E$.
    (c) If $\Pr(E \mid C, K_i) \leqslant \Pr(E \mid \overline{C}, K_i)$ for all $i$, then $C$ is a *Pareto negative* causal factor for $E$.
    (d) If $\Pr(E \mid C, K_i) < \Pr(E \mid \overline{C}, K_i)$ for all $i$, then $C$ is a *negative* causal factor for $E$.
    (e) If $\Pr(E \mid C, K_i) > \Pr(E \mid \overline{C}, K_i)$ for some $i$, and $\Pr(E \mid C, K_i) < \Pr(E \mid \overline{C}, K_i)$ for some $i$, then $C$ is a *mixed* causal factor for $E$.

15. Note that this theory is explicitly and intentionally non-reductionist. The definition we provided of *causal background context* required us to say what the other causes of $E$, excepting $C$, were. So it does not allow us to construct the causal relations from the probability function alone. Rather, what the account ends up providing us is something about the *relationship* between probability and causation.

16. Consider how Eells's account handles the *type* of indeterministic neuron diagram shown in figure 4.5 (built to mimic Hesslow (1976)'s birth control case).

    (a) In this neuron diagram, the numbers outside of parentheses give the probability (along that route) of the neuron at the head of the arrow firing, given that the neuron at the base of the arrow has fired, and the numbers inside the parentheses give the probability (along that route) of the neuron at the head of the arrow firing, given that the neuron at the base of the arrow has *not* fired.

        i. The idea here is that we can think of each arrow between the neurons as being either a 'matching' path or a 'non-matching' path. If it is a matching

path, then the firing of the neuron at the base of the arrow will send a stimulatory signal to the neuron at the head of the arrow. If it is a non-matching path, then the *non-firing* of the neuron at the base of the arrow will send a stimulatory signal to the neuron at the head of the arrow. The number outside parentheses along each path is the probability that it is a matching path, conditional on the neuron at the base firing; and the number inside parentheses is the probability that it is a matching path, conditional on the neuron at the base *not* firing.[6]

(b) There are no causes of $T$ which are *independent* of $B$. So there is only the trivial causal background context. And, as you can check for yourself,

$$\Pr(T \mid B) = \frac{19}{25} = 0.76 = \Pr(T \mid \overline{B})$$

So $B$ is not a type causal factor for $T$, according to Eells's theory.

(c) Yet, it seems that, on one reading, the type causal claim "$B$ firings cause $T$ firings" is true.

i. Of course, there's another sense in which the type causal claim "$B$ firings *prevent* $T$ firings" is true (they do so by preventing $P$ firings). And (perhaps) there's a further sense in which it's true to say that $B$ firings don't have any net effect on $T$ firings. The object to Eells is that his theory of type causation doesn't allow any sense in which "$B$ firings cause $T$ firings" is true.

(d) Eells's response appeals to the hypothetical frequentist interpretation of probability. On that theory, probabilistic claims are made relative to a particular *kind*.

(e) While it's true that, relative to the kind *trials of this indeterministic neuron system*, $B$ is causally neutral for $T$, relative to the kind *trials of this indeterministic neuron system in which P fires*, as well as relative to the kind *trails of this indeterministic neuron system in which P doesn't fire*, $B$ is causally positive for $T$'s firing.

17. Eells (1991) additionally provides a measure of the causal influence of $C$ on $E$, $\mathfrak{I}(C, E)$; this is given by the probability increase of $E$, conditioned on $C$, within each background context $K_i$, weighted by the probability of that background context.

$$\mathfrak{I}(C, E) = \sum_i \Pr(K_i) \cdot \left[\Pr(E \mid CK_i) - \Pr(E \mid \overline{C}K_i)\right]$$

[6] And, as before, the probabilities along the various routes are independent, so that the probability that $T$ fires, given that *both* $B$ and $P$ fire, is just the probability that either the $B \to T$ path or the $P \to T$ path is matching, conditional on both $B$ and $P$ firing—*i.e.*,

$$\frac{3}{5} + \frac{2}{3} - \frac{3}{5} \cdot \frac{2}{3} = \frac{13}{15}$$

(a) Note that, if we had instead written

$$\sum_i \Pr(E \mid CK_i) \cdot \Pr(K_i \mid C) - \Pr(E \mid \overline{C}K_i) \cdot \Pr(K_i \mid \overline{C})$$

this would be equivalent to $\Pr(E \mid C) - \Pr(E \mid \overline{C})$. The action in this measure of causal strength lies, then, in the way that it *factors out* the evidential relevance of $C$ to the causal background contexts by replacing $\Pr(K_i \mid C)$ and $\Pr(K_i \mid \overline{C})$ with $\Pr(K_i)$.

### 4.5.2 Token Causation

18. Eells's theory of token causation presupposes a rather coarse-grained theory of events, according to which a token event $e$ can be of several different types. Importantly, according to this theory, whether the token event $c$ caused $e$ can depend upon which types of $C$ and $E$ we are considering. The relation to be explicated is one of '$c$'s exemplifying type $C$ token causing $e$'s exemplifying type $E$'. Thus, even though Eells does not think that *events* are fine-grained, he does think that the causal relata are fine-grained.

19. Eell's theory of token causation involves three different notions: that of $e$'s exemplifying $E$ *because of*, *despite*, and *independently of* $c$'s exemplifying $C$. Only the first of these is genuine token causation, and so we'll focus on that notion here.

20. The token account also relies upon the notion of a *causal background context*. However, the notion is different in the case of token causation. In the case of token causation, the relevant causal background context (for discovering whether $c$'s exemplifying $C$'s caused $e$'s exemplifying $E$) is determined as follows:

    (a) Take all factors $F_1, F_2, \ldots, F_N$ which are *not token caused* by $c$'s exemplifying $C$ but which *are* type causally relevant to $e$'s exemplifying $E$ (exluding $C$, if $C$ is indeed type-causally relevant to $E$).[7]

    (b) Let '$F_i^{@}$' be the actual value of these factors, on the occasion in question.

    (c) Then, the causal background context is $K_{@} = \bigwedge_i F_i^{@}$.

21. The first-pass version of Eell's theory of token causation is this:

> Eells's Theory of Token Causation (v1)
> The token event $c$'s exemplifying type $C$ at $t_c$ caused the token event $e$'s exemplifying type $E$ at $t_e$ iff:[8]
> (a) $t_c < t_e$;
> (b) $\Pr_{>t_c}(E \mid K_{@}) > \Pr_{<t_c}(E \mid K_{@})$;

[7] This is tweaked in §6.3, for reasons similar to the ones that the type account was tweaked in chapter 3, but we'll ignore those complications here.

[8] We are assuming, as usual, that both $c$'s exemplifying type $C$ at $t_c$ and $e$'s exemplifying type $E$ at $t_e$ actually occur.

(c) $\Pr_{>t_c}(E \mid K_@)$ is high; and

(d) $\Pr_t(E \mid K_@)$ remains high for all $t < t_e$.

(a) Here, '$\Pr_{>t_c}$' is the probability function at a time *just after* $t_c$, and '$\Pr_{<t_c}$' is the probability function at a time *just before* $t_c$.

(b) This probability changes over time by conditionalizing on the values of the factors which are type-causally relevant to $E$, as those factors assume values.

22. Eells provides us with a definition of the *degree* $\delta$ to which a token event $c$'s exemplifying $C$ caused the token event $e$'s exemplifying $E$:

$$\delta(c, C, e, E) = \min\{\Pr_t(E \mid K_@) \mid t_c < t < t_e\} - \Pr_{<t_c}(E \mid K_@)$$

Or, if we just write '$\delta_{c,e}$' for '$\delta(c, C, e, E)$', '$M$' for the minimum value of $\Pr_t(E \mid K_@)$, for $t$ between $t_c$ and $t_e$, and '$P$' for the prior probability of $E$, before $t_c$, then

$$\delta_{c,e} = M - P$$

23. With this notion of degree of causation, Eells then refines the account to take into account not only the factors that *c doesn't* token cause, but additionally, those which it *does* token cause, but where $c$ is not the *only* token cause. His idea is to somehow *factor out* the contribution made by $c$ to these intermediate factors.

(a) Let $K_@$ be constructed as before. Then, consider, *in addition*, the event of $f$'s being $F$ which is, to some degree, token caused by $c$'s being $C$.

(b) We may define $\Pr_t^f$ as follows:

$$\Pr_t^f(-) \stackrel{\text{def}}{=} (1 - \delta_{c,f}) \Pr_t(- \mid F) + \delta_{c,f} \Pr_t(- \mid \overline{F})$$

(c) Eells then uses *this* probability function to give his account of token causation. The account is the same as before, except that we should add the superscript '$f$' for the factor which $c$'s being $C$ token causes to some degree short of 1.

i. How should we define this function for *multiple* factors which $c$'s being $C$ token caused? Eells: "iteratively" (p. 362). I'm open to hearing suggestions about what this means.

Eells's Theory of Token Causation (v2)
The token event $c$'s exemplifying $C$ at $t_c$ caused the token event $e$ exemplifying $E$ at $t_e$ iff:

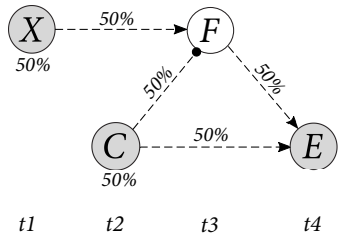
(a) $t_c < t_e$;

(b) $\Pr_{>t_c}^f(E \mid K_@) > \Pr_{<t_c}(E \mid K_@)$;

(c) $\Pr_{>t_c}^f(E \mid K_@)$ is high; and

(d) $\Pr_t^f(E \mid K_@)$ remains high for all $t < t_e$.

**Figure 4.6**




## 4.6 Objections to Eells' Theory

24. A nice place to start, in general, when considering theories of causation is with cases of preemption. So let's just consider the indeterministic neuron diagram shown in figure 4.6. Let's begin by looking at what happens with the initial version of the theory.

    (a) In this case, the causal background context for the influence of $C$'s firing at $t_2$ on $E$'s firing at $t_3$ includes only $X$'s firing at $t_1$. If we look at the probability distribution over $C$, $E$, and $F$ *conditionalized* on the fact that $X$ fired, we get the following:

| | $\Pr(- \mid X)$ |
|---|---|
| $CFE$ | $3/32$ |
| $CF\overline{E}$ | $1/32$ |
| $C\overline{F}E$ | $6/32$ |
| $C\overline{F}\,\overline{E}$ | $6/32$ |
| $\overline{C}FE$ | $4/32$ |
| $\overline{C}F\overline{E}$ | $4/32$ |
| $\overline{C}\,\overline{F}E$ | $0$ |
| $\overline{C}\,\overline{F}\,\overline{E}$ | $8/32$ |

    (b) Then, as you can verify for yourself,

$$\Pr_{t1}(E) = \frac{13}{32}$$
$$\Pr_{t2}(E) = \frac{18}{32}$$
$$\Pr_{t3}(E) = \frac{16}{32}$$

25. So, if we assume that any value above $1/2$ counts as 'high', then $C$'s firing caused $E$'s firing, according to the first version of the account.

26. However, notice that the *way* that it did this was slightly odd. It really mattered how high the probability was at $t_3$. Suppose that we tweak the indeterministic neuron

diagram from figure 4.6, so that the probability that the path between $C$ and $E$ is active is 25%, rather than 50%. Then, we'll get the following probability distribution, conditionalized on $X$:

| | $\Pr(- \mid X)$ |
|---|---|
| $CFE$ | $5/64$ |
| $CF\overline{E}$ | $3/64$ |
| $C\overline{F}E$ | $6/64$ |
| $C\overline{F}\overline{E}$ | $18/64$ |
| $\overline{C}FE$ | $8/64$ |
| $\overline{C}F\overline{E}$ | $8/64$ |
| $\overline{C}\overline{F}E$ | $0$ |
| $\overline{C}\overline{F}\overline{E}$ | $16/64$ |

(a) Then, we'll have the following probability trajectory:

$$\Pr_{t1}(E) = \frac{29}{64}$$
$$\Pr_{t2}(E) = \frac{22}{64}$$
$$\Pr_{t3}(E) = \frac{16}{64}$$

(b) And $C$'s firing at $t_2$ will *not* count as a token probabilistic cause of $E$'s firing at $t_3$.

(c) This looks like a bad result. $C$'s firing at $t_2$ *was* the token cause of $E$'s firing at $t_3$. It was the *only* token cause of that event.

27. Perhaps we need to wheel in the complications of the second version of the account. To do so, we'll first have to figure out the value of $C$'s degree of influence on $F$'s failure to fire, $\delta_{c,\overline{f}}$.

(a) First, look at the probability trajectory of $F$'s probability of firing, holding fixed $X$:

$$\Pr_{t1}(\overline{F}) = \frac{39}{64}$$
$$\Pr_{t2}(\overline{F}) = \frac{46}{64}$$

So, $\delta_{c,\overline{f}} = 7/64$.

28. Now, we can calculate $\Pr^f$. It is just

$$\Pr^f(-) = (57/64) \cdot \Pr(- \mid \overline{F}) + (7/64) \cdot \Pr(- \mid F)$$

(a) Since

$$\Pr_{t1}(E \mid F) = \frac{13}{24} \qquad \Pr_{t1}(E \mid \overline{F}) = \frac{3}{20}$$
$$\Pr_{t2}(E \mid F) = \frac{15}{24} \qquad \Pr_{t2}(E \mid \overline{F}) = \frac{1}{4}$$
$$\Pr_{t3}(E \mid F) = \text{undefined} \qquad \Pr_{t3}(E \mid \overline{F}) = \frac{1}{4}$$

(b) We'll have the following probability trajectory (assuming that we treat the undefined probability value as zero):

$$\Pr^{f}_{t1}(E) \approx 0.499$$
$$\Pr^{f}_{t2}(E) \approx 0.584$$
$$\Pr^{f}_{t3}(E) \approx 0.027$$

(c) And, again, $C$'s firing will not be a token cause of $E$'s firing.

# 5 Process Theories

1. The basic idea behind process theories is that the token event $c$ caused the token event $e$ if and only if we can trace out the right kinds of *causal processes* leading from $c$ to $e$.

   > Process Theory
   > A token event $c$ caused a token event $e$ if and only if $c$ and $e$ are connected by a series of intersecting *causal processes* whose intersections constitute *causal interactions*.

   (a) In fact, Dowe (2000) does not accept both directions of the biconditional above; but it helps to get a preliminary understanding of what's going on with process theories.

   (b) Note that this is a theory of *token* causation; *type* causal claims are understood as generic claims about token-level causal relations.

2. According to a process theory, the binary relation of *causation* is not the fundamental causal notion. Fundamentally, causality is a matter of causal processes. When we speak of causal *relations*—*i.e.*, when we say things like 'Suzy's throw caused the window to shatter', what makes such claims true is the existence of the right kind of causal process connecting Suzy's throw with the window's shattering.

3. A process theory owes us an account of the two central notions appearing in this definition—that of a *causal process* and that of a *causal interaction*.

   (a) Though Salmon (1984) appeals to the notion of *mark transmission* from Reichenbach (1956), both Salmon (1994) and Dowe (2000) give accounts of these notions which appeal to quantities which are *conserved* at the actual world.

   (b) Therefore, as we saw when we discussed Dowe's methodology, a process theory will be at best only true at physically possible worlds.

## 5.1 Salmon's Mark Transmission Theory

1. Salmon begins with the notion of a *process*. While no precise definition of process is given, we can perhaps get by with the following understanding: a process is just the 4-dimensional worm of a persisting entity.

(a) Sacrificing rigor, let's refer to one of these 4-D worms as the 'world-line' of the entity.

2. Some processes are *pseudo processes*, whereas some processes are *causal processes*. For instance:

   (a) The world-line of a beam of light traveling from the spotlight to the wall is a *causal process*; the world-line of the spot of light shining on the wall as it moves around is a *pseudo-process*.

   (b) The world-line of the car is a *causal process*; the world-line of the car's shadow is a *pseudo-process*.

3. One reason Salmon (1984) wishes to distinguish pseudo- from causal-processes is that he takes special relativity to say that no causal process may travel faster than the speed of light. But a spot of light or a shadow *may* travel faster than the speed of light. Such processes must not be considered causal, then.

4. A causal process is defined in terms of its ability to transmit marks.

   (a) A *mark* is defined simply as the change of some property of the process. The process is marked iff it begins with some property $Q$, and comes, after some causal interaction, to adopts have an incompatible property $Q'$.

   (b) To *transmit* a mark is simply to *have* the mark at every point in spacetime along its worldline.

5. More carefully:

   Causal Process
   A process $P$ is *causal* iff there is some property $Q$ such that:
   (a) in the absence of *causal interactions* with other processes, $P$ would manifest the property $Q$ at every point in some interval of spacetime points;
   (b) there is some possible *causal interaction* $P$ could undergo such that, were $P$ to undergo this causal interaction,
      i. after the interaction, $P$ would not manifest $Q$, but rather an incompatible property $Q'$; and
      ii. in the absence of further interventions, $P$ would continue to manifest $Q'$.
   A process is a *pseudo-process* iff it is not causal.

   (a) Note that this definition involves three counterfactual conditionals. One of these—(b)i—says that the process would be marked by the appropriate causal interaction. The other two stipulate that the mark would not be transmitted in the absence of the interaction—(a)—and that it would be transmitted *without further interventions*—(b)ii. To make this clearer, we could re-write the account of causal process as follows:

Causal Process
A process $P$ is causal iff there is some property of $P$, $Q$, and some potential causal interaction $I$ such that

i. were $P$ to undergo the interaction $I$, it would transmit the mark $Q'$ (where this is an alteration of $Q$);
ii. were there to be *only* the interaction $I$, and no other interventions, $P$ would still transmit that mark; and
iii. were there no causal interactions with $P$, $P$ would not transmit that mark.

(b) To motivate these counterfactual criteria, Salmon notes the following example, due to Nancy Cartwright:

i. The moving spot of light comes across a red surface. When it crosses the red surface, a red filter is put in front of the spotlight. The spot of light remains red thereafter.
ii. Note that the red filter at the bulb is not a causal interaction *with the process of the spot of light*. So it is important that we say "no other *interventions*", and not "no other causal interactions with $P$".

(c) Note also that the definition involves reference to the idea of a *causal interaction*, which we have yet to define.

i. The notion of a *causal interaction* is therefore (definitionally) more primitive than the notion of a *causal process*.

6. Salmon (1984) gives the following definition of a *causal interaction*.

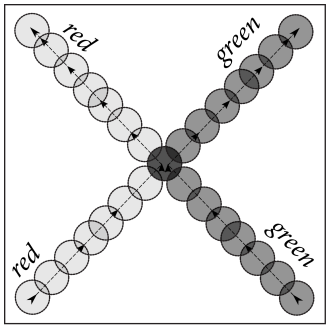
Causal Interaction
A causal interaction is the intersection of two processes $P_1$ and $P_2$ such that:

(a) $P_1$ manifests the property $Q$ prior to the interaction and the incompatible $Q'$ after the interaction;
(b) $P_2$ manifests the property $R$ prior to the interaction and the incompatible $R'$ after the interaction;
(c) were $P_1$ to not intersect $P_2$, $P_1$ would continue to manifest $Q$; and
(d) were $P_2$ to not intersect $P_1$, $P_2$ would continue to manifest $R$.

(a) This definition, too, involves counterfactual conditionals. To motivate these counterfactual conditionals, Salmon asks us to consider the following case:

i. A red spot of light moves from the lower left corner of a screen to its center as a green spot of light moves from the lower right corner of a screen to its center. They meet at the center, and then the red spot of light moves towards the upper left corner, while the green spot of light moves towards the upper right corner. (see figure 5.1).
ii. The are two processes which begin with the red spot of light's motion: the one which remains red after the intersection and heads back towards the upper left corner, and the one which becomes green after the intersection

**Figure 5.1** Salmon's 'colliding spots of lights'




and heads towards the upper right corner. Call the first of these the 'red-green' process, and the second of these the 'red-red' process. Similarly, we can distinguish the 'green-red' process and the 'green-green' process.

iii. It *appears* as though the red-green process interacted with the green-red one (both of them changed their color after the intersection). Similarly, it *appears* as though the red-red process interacted with the green-green one (both of them changed their velocity after the intersection). However, Salmon does not wish to call these *genuine* interactions.

iv. The counterfactual formulation of the definition of causal interaction is meant to rule this out. For the following counterfactuals are false:

   A. Had the red-green process not intersected green-red process, it would have maintained its redness.

   B. Had the red-red process not intersected the green-green process, it would have maintained its velocity.

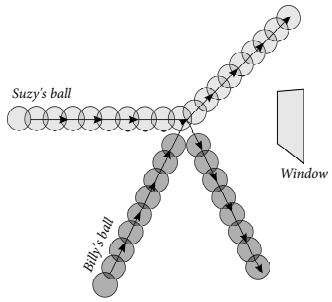
   So, by clauses (c) and (d) of the definition of 'causal interaction', this does not constitute a causal interaction.

## 5.2 Objections to Salmon's Theory

7. As stated, Salmon's theory does not say anything about which *kind* of properties $Q$ are the kind which get to enter into his definitions of causal interaction and causal process. One objection to the theory points out that, if any old properties are allowed to enter into this theory, then shadows will end up counting as causal processes after all.

   (a) Suppose that a shadow is cast by a flagpole and that $x$ meters from the base of the flagpole is a crack in the pavement. Suppose that the shadow moves from an original distance $2x$ from the base of the flagpole to a distance $x/2$ from the base of the flagpole.

**Figure 5.2** Prevention




(b) Let $Q$ be the property of being at least distance $x$ from the base of the flagpole, and let $R$ be the property of being dark.

(c) Then, the process of the shadow intersects the process of the crack in the pavement; and this intersection seems to count as a causal interaction according to the definition given above, for:

i. prior to the intersection, the shadow is $Q$ and after the intersection, the shadow is $\neg Q$;

ii. prior to the intersection, the crack in the pavement is $R$ and after the intersection, the crack is $\neg R$;

iii. had the processes not intersected, the shadow would not have been $\neg Q$;

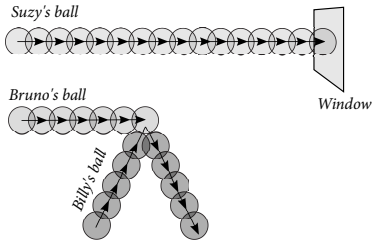
iv. had the processes not intersected, the crack would not have been $\neg R$.

8. To solve this problem, Salmon must say something about which kinds of properties are appropriate.

9. A standard objection to process theories of causation concerns cases of *prevention* and *double prevention*.

> Prevention
> Suzy's ball is on a path to hit the window. Billy, attempting to save the window, throws his ball at just the right time with just the right momentum to deflect Suzy's ball. They collide, Suzy's ball is deflected, and the window remains intact. (See figure 5.2.)
>
> Double Prevention
> Suzy's ball is on a path to hit the window. Billy, attempting to save the window, throws his ball at just the right time with just the right momentum to deflect Suzy's ball. Bruno, seeing Billy's plan, throws his ball at just the right time and with just the right momentum to deflect Billy's ball so that it doesn't deflect Suzy's ball. Billy's ball is deflected, and Suzy's ball shatters the window. (See figure 5.3.)

**Figure 5.3** Double Prevention




(a) In Prevention, Billy's throw saved the window (*i.e.*, it caused the window to remain unshattered). However, there is no intersection of the process of Billy's ball and the process of the window. So there can be no causal interaction between them, on Salmon's theory; consequently, they cannot be causally related.

(b) In Double Prevention, Bruno's throw caused the window to shatter. However, there is no intersection of the process of Bruno's ball and the process of the window. So there can be no causal interaction between them, on Salmon's theory; consequently, they cannot be causally related.

i. Why mention both Prevention and Double Prevention?

A. Because Dowe (2000, p. 123) claims, in his treatment of these issues, that these kinds of cases involve omissions, or negative events, "as one or both of the relata"; and the fact that these relations involve negative events strengthens his contention that "we can recognize that it is not literal causation".

B. He therefore understands the problem posed by cases of prevention as the problem of "how negative events can enter into real causal relations."

C. I believe that cases of double prevention show that this is mistaken. Both Bruno's throw and the window's shattering are positive events, and those positive events are causally related. The problem posed by prevention is not merely a problem of incorporating negative events.

## 5.3 Dowe's Conserved Quantity Theory

10. Dowe's theory makes use of the same general idea as Salmon's: causal processes are fundamental, and events are causally related iff they are connected by a series of causal processes whose intersections constitute causal interactions.

11. However, he does not understand either *causal process* or *causal interaction* in the way that Salmon does.

(a) For Dowe, a *process* is the worldline of an *object*.

i. Not just any thing you can name will count as an *object*, on Dowe's view; therefore, not just any 4-dimensional worm will count as a process, on Dowe's view.

A. Therefore, Dowe divides worldlines into those which are *causal processes*, those which are *pseudo processes*, and those which are *spatio-temporal junk* (Kitcher (1989)'s term).

ii. Objects are "anything found in the ontology of science...or common sense" (p. 91).

(b) A process is *causal* iff it is the worldline of an object which possesses a conserved quantity.

> Causal Process
> A *causal process* is the worldline of an object which possesses a conserved quantity.

i. A quantity is *conserved* iff the global amount of that quantity must remain fixed, as a matter of the laws of physics.

A. Which quantities are conserved is a matter to be decided by physics, not philosophy. However, it seems plausible that charge, mass-energy, and linear momentum as plausible candidates.

B. Why 'global'? Why not say that a conserved quantity is one which must, as a matter of the laws of physics, remain fixed within any closed system? Because a system is only closed if it is free from external causal influence; therefore, the notion of a 'closed system' is a causal one. Dowe wishes to avoid circularity, so it is important to his project that the quantities be conserved *globally*.

C. In this connection: Noether's Theorem establishes a connection between conservation laws and symmetries of space-time; and there are spacetimes consistent with general relativity in which mass-energy is not globally conserved. (My sources in cosmology tell me that, in fact, mass-energy is *not* globally conserved. I'd be interested in hearing more from those who understand the physics better than I do.)

(c) An intersection of two causal processes constitutes a *causal interaction* iff, at the point of intersection, there is an *exchange* of a conserved quantity.

> Causal Interaction
> The intersection of two causal processeses, $P_1$ and $P_2$, constitutes a *causal interaction* iff there is some conserved quantity $Q$ such that:
>
> i. prior to the intersection, the object in $P_1$ had quantity $q_1$ of $Q$;
> ii. prior to the intersection, the object in $P_2$ had quantity $q_2$ of $Q$;
> iii. after the intersection, the object in $P_1$ had quantity $q_1' \neq q_1$ of $Q$; and
> iv. after the intersection, the object in $P_2$ had quantity $q_2' \neq q_2$ of $Q$.

12. Note that Dowe's theory, unlike Salmon's, does not appeal to any counterfactuals.

(a) This is in part because Dowe does not characterize causal processes in terms of *marks*, and so does not face the kinds of complications which Salmon's mark transmission theory did.

13. Consider a spot of light moving around a screen. This spot of light counts as an object according to common sense, so the world line of the spotlight counts as a process.

    (a) Why not a causal process? Dowe: because the spot of light does not possess any conserved quantities, though the screen it is projected onto will possess conserved quantities.

        i. How can we tell whether an object 'possesses' a conserved quantity or not? (Dowe, 2000, p. 92): "An object possesses energy if science attributes that quantity to that body."

    (b) Salmon (1994): very well, then let us not consider the world line of the spot of light, but rather the world line which is the region of the screen illuminated by that light. The screen may possess conserved quantities, so the definition of Causal Process will rule that the region of the screen illuminated by the spot of light is a causal process. But this process can travel faster than the speed of light (since the spot of light can).

        i. (Dowe, 2000, p. 99): the region of the screen illuminated by the spot of light is not an object. So there is no process here; merely spatio-temporal junk.

14. As yet, we have done nothing to address the worries about cases of prevention and double prevention.

    (a) Dowe does not deal with this problem by complicating the account of causation provided above; rather, he first distinguishes such cases from causation proper by calling them cases of 'causation*'.

    (b) He then claims that "causation* should be understood not as real causation but as a hybrid fact usually involving certain actual real causation together with certain counterfactual truths about real causation...nevertheless, we are justified in treating such cases as causation for practical purposes" (p. 124).

    (c) That causation* is not real causation is argued for by presenting the following dialogue, which is meant to bring out the "intuition of a difference" (p. 125):

        > ...when you talk of a scenario in which the father presented the accident, you don't mean that he bore any literal causal connection to a real thing called an accident. You mean that had he not acted in the way that he did, some circumstances would have brought about the accident.
        > Yes, that's exactly what I mean.

    (d) Dowe says that most cases of prevention will involve a negative event as the effect. He therefore gives the following analysis of prevention in which the effect is a negative event:

Causation* by Prevention
*C* prevented *E*—that is, *C* caused* not-*E* if *C* occurred and *E* did not, and there was an event *X* such that:

i. there is a causal relation between *C* and the process *P*, of the object *x*, such that either:
   A. *C* is a causal interaction with *P*; or
   B. *C* causes *Y*, which is a causal interaction with *P*; and
ii. If *C* had not occurred, then *x* would have caused *E*.

(e) This tells us, correctly, that Billy's throw caused (or caused*) the window to not shatter.

(f) However, it does not tell us that Bruno's throw caused (or caused*) the window to shatter—though we could complicate the account with further counterfactuals to fix that.

## 5.4 Objections to Dowe's Theory

15. I think there are two principal objections to Dowe's theory: first, that it rules out clear cases of causation; and secondly, that it lets in clear cases of non-causation. Let's take these in order.

16. First: the account rules out clear causes of causation.

    (a) Jonathan Schaffer (2000) responds to Dowe's stance on prevention by pointing out that many paradigm cases of causation are actually, on a more detailed analysis, cases of double prevention—that is, on Dowe (2000)'s analysis, they are really cases of 'causation*'.

    (b) I push the detonator, and the bomb explodes. Suppose that the mechanism of the detonator works like this: there is an inhibitor of the bomb's detonation which requires electricity in order to do its work. Pressing the button cuts the electricity; with the electricity cut, the inhibitor fails, and the bomb explodes.

    (c) Then, we just have a case of double prevention. But, in cases of double prevention, there is no causal process leading from the cause to the effect.

    (d) Surely this is causation *full stop*. There is no 'intuition of a difference' here.

17. Secondly: the account lets in clear cases of non-causation.

    (a) The light emits photons. Those photons rebound off of the wall. After this intersection with the worldline of the wall, they have a different momentum. Thus, the intersection constitutes a causal interaction. These photons then rebound off of the man, who proceeds to dance. This again is a causal interaction.

        i. First problem: the wall did not cause the man to dance. But, there is a series of causal processes whose intersections constitute causal interactions connecting them.

ii. Second problem: there are two ways we could fill out this case:
    A. The man was waiting for the light to be turned on before he started dancing. In this case, the light caused the man to dance.
    B. The man was counting to ten, and was going to dance when he got to ten, whether the lights were on or not. In this case, the light did not cause the man to dance.

(b) Dowe responds to these kinds of cases by emending his account. He *denies* that $c$ causes $e$ iff they are connected by a series of causal processes whose intersections are causal interactions. Rather, he says that this is merely a *necessary*, but not a *sufficient* condition for $c$ causing $e$.

(c) The revised account supposes that the causal relata are facts which *supervene* upon the possession of conserved quantities. So we can characterize the causal relata in terms of their possession of conserved quantities.

i. Dowe writes out the fundamental causal relata as '$q(e) = x$', which represents the fact that the object $c$ possesses $x$ amount of conserved quantity $q$ at the relevant time.
ii. He claims that all causal relata will supervene upon fundamental causal relata of this form.

Then, (Dowe, 2000, p. 171) gives the following account of causation:

> Causal Connection
> There is a causal connection between a fact $q(c) = x$ and $q'(e) = y$ iff there is a series of causal processes and interactions between $q(c) = x$ and $q'(e) = y$ such that:
> i. any change of object from $a$ to $b$ and any change of conserved quantity $q$ to $q'$ along this path occur at a causal interaction involving the following changes: $\Delta q(a)$, $\Delta q(b)$, $\Delta q'(a)$, and $\Delta q'(b)$; and
> ii. for any exchange in (i) involving more than one conserved quantity, the changes in quantities are governed by a single law of nature.

(d) Now, the man's dancing is supposed to supervene upon the possession of some conserved quantities $q(e)$, and the light's being illuminated is supposed to supervene upon the possession of some conserved quantities $q(c)$; and the light's being illuminated is supposed to cause the man to dance iff there is a causal relation between the fundamental facts about the possession of the conserved quantities $q(c)$ and $q(e)$ upon which those events supervene.

18. I have a very hard time understanding how this account is going to distinguish the case in (17(a)iiA) from the case in (17(a)iiB).

(a) It seems that whatever conserved quantities the light's being illuminated supervenes upon in the first case, those very same conserved quantities will be possessed in the second; and it seems as thought, whatever conserved quantities the man's dancing supervenes upon in the second cacse, those very same conserved quantities will be possessed in the second.

(b) So it seems that the account doesn't have the resources to distinguish the two cases (but I must admit that I'm very unclear on how the account is supposed to work).

(c) Here is what (Dowe, 2000, p. 174) has to say about a similar case involving a man tapping on the desk and the causal process of sound waves emanating from the tapping, which reverberate a ticking clock hand (though, intuitively, we don't want to say that the man tapping causes the clock hand to tick):

> The alleged effect, the hand moving, no doubt supervenes on a genuine physical fact involving a conserved quantity. But it is not connected to the sound waves, since whatever quantitiy is involved in the string of processes and interactions that constitute the sound waves, it is not the quantity that changes in the interaction underlying the 'hand moving'. If, perchance, the motion of air molecules did make some slight difference to a distant event, then that would qualify as a (minor partial) cause.

# 6 Counterfactual Theories

## 6.1 Primer on 'Counterfactuals'

1. What philosophers (somewhat inaptly) call 'counterfactuals' are *conditional* sentences (sentences of the form ⌜if $A$, then $C$⌝) in the *subjunctive* mood.

   (a) The 'if' part of a conditional is called its 'antecedent'; the 'then' part is called its 'consequent'.

   They are contrasted with the more aptly named 'indicative conditionals'.

2. The following pairs of conditionals are often used to illustrate the distinction between indicative and subjunctive conditionals (in both cases, we assume that the conspiracy theories are false).

   (a) i. If Oswald didn't shoot Kennedy, then somebody else did.
   ii. If Oswald hadn't shot Kennedy, then somebody else would have.

   (b) i. If Shakespeare didn't write Hamlet, then somebody else did.
   ii. If Shakespeare hadn't written Hamlet, then somebody else would have.

   (c) While (2(a)i) and (2(b)i) (the indicative conditionals) are true, (2(a)ii) and (2(b)ii) (the counterfactuals) are false.

3. Subjunctive conditionals are called 'counterfactuals' because the original namers thought that they always presuppose the falsehood of their antecedents.

   (a) Not so—witness: 'if he had the measles, he'd be showing all the symptoms he in fact is showing.'

4. The first systematic (and still paradigm) semantics for counterfactuals came from Lewis (1973b) and Stalnaker (1968). The basic idea behind those theories (papering over several complications) is the following.

   (a) We have a *selection function*, $f$. $f$ is a function from pairs of propositions and possible worlds to *sets* of possible worlds. The interpretation is that $f(A, w)$ is the set of $A$-worlds which are *most similar* to $w$.[1]

[1] An '$A$-world' is just a possible world at which $A$ is true.

(b) A counterfactual ⌜if it were the case that $A$, then it would be the case that $C$⌝ (symbolized by Lewis with ⌜$A \mathrel{\Box\!\!\rightarrow} C$⌝) is true at a world $w$ iff all of the $A$-worlds most similar to $w$ are $C$-worlds.

$$A \mathrel{\Box\!\!\rightarrow} C \text{ is true at } w \iff f(A, w) \models C$$

(c) Lewis assumes that, if $w$ is an $A$-world, then $f(A, w) = \{w\}$. This assumption is known as 'strong centering', and, given his semantics, is equivalent to the claim that '$A \mathrel{\Box\!\!\rightarrow} A$' is necessarily true for all $A$.

   i. The condition could be weakened to 'weak centering', which is the claim that, if $w$ is an $A$-world, then $w \in f(A, w)$. Given Lewis's semantics, this is equivalent to the claim that modus ponens is valid for the counterfactual conditional.

(d) For our purposes here, we can lean on the understanding of $f$ that we get from Maudlin (2007) (which accomplishes by fiat the thing that that Lewis (1979) tried to accomplish with explicit standards of similarity): $f(A, w)$ is the set of worlds that you get by, just prior to the antecedent time, minimally altering $w$ so as to make $A$ true, and then time-evolving the resulting state forward in time according to the fundamental equations of motion.

   i. This account of the selection function has the effect of building in the temporal direction of causation. Lewis wanted to have this be a *consequence* of some standards of similarity, so that the direction of causation would end up being a consequence of the fact that entropy increases into the future and decreases into the past at the actual world.

   ii. Elga (2001) argues persuasively that Lewis (1979)'s standards of similarity fail in this goal, so I'm putting these complications to the side.

(e) We should be clear that we are ruling out, from the get-go, the so-called 'backtracking' reading of counterfactuals. That is, you can hear both of the sentences as being true:

   i. If John had jumped, he would have died (since he wasn't wearing a parachute).

   ii. If John had jumped, he would have lived (since, had he jumped, he would have to have been wearing had a parachute).

   The second of these counterfactuals is a 'backtracker'. It considers not just what would follow from the counterfactual assumption of the antecedent, but also what the necessary preconditions for the truth of the antecedent would have to be. The counterfactuals we are concerned with here will all be 'non-backtracking' counterfactuals.

### 6.2 Lewis's 1973 Counterfactual Theory of Causation

5. Say that one event, $e$, *causally depends upon* a distinct event, $c$, iff, had $c$ occurred, $e$ would have occurred; and, had $c$ not occurred, $e$ wouldn't have occurred either.

Causal Dependence (v1)
The event $e$ causally depends upon the event $c$ iff:

$$O(c) \mathrel{\Box\!\!\rightarrow} O(e) \wedge \neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$$

(a) '$O(c)$' is the proposition that the event $c$ occurs; '$O(e)$' is the proposition that the event $e$ occurs.

(b) Assuming strong centering, if $c$ and $e$ are actually occurring events, then the first conjunct of the above will be trivially satisfied, so we can simply say:

Causal Dependence (v2)
The occurrent event $e$ causally depends upon the occurrent event $c$ iff:

$$\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$$

6. On Lewis (1973a)'s view, causal dependence between distinct events is sufficient for causation.

(a) Note that, given the second definition above, this means that

i. Caesar's birth caused his death.

is true, on this account. For, had Caesar not been born, his actual death would not have occurred.

i. An odd consequence—but Lewis (2004, p. 101) is not bothered.

ii. Note, however, that while Lewis is not bothered by it, this consequence not a necessary one. If we weaken strong centering to weak centering, then we could insist that, even though the set of most similar worlds at which Caesar is born *includes* the actual world, it includes *other* worlds as well. And in some of these worlds, Caesar does not die the death he actually died. (The world is chaotic; tiny variations in the manner of Caesar's birth are sufficient to send his life off on a different trajectory altogether—he would still have died, but he would not have died the death he actually died.)

(b) Note that, here, "distinct" means more than just "not identical".

i. Suppose that John plays Poker. Then, he also plays cards (*by* playing Poker). And, given Lewis's theory of events (1986b), his playing cards is not identical to his playing Poker (the latter is more *modally fragile* than the former; its otherworldly regions are a subset of those of the former).

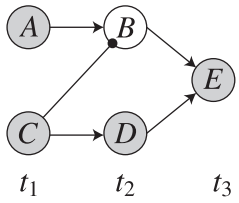
ii. And the following counterfactual is true: "Had John not played cards, he would not have played Poker".

iii. So, if "distinct" just meant "non-identical", then we would have to conclude that John's playing cards caused him to play Poker.

iv. Lewis thus understands "distinct" in such a way that $e_1$ and $e_2$ are not distinct if they are identical, but also if they are nonidentical but one *implies* the other (one event *implies* another just in case the occurrence of the one entails the occurrence of the other), or the two events *overlap* (there is another event that both the events have as a part).[2]

[2] See Lewis (1986b).

**Figure 6.1** A case of preemption




7. Though causal dependence is sufficient for causation, it is not necessary. Cases of PREEMPTION suffice to demonstrate this (see figure 6.1).

   (a) If we minimally alter the world so that $C$'s firing at $t_1$ does not occur, then $A$'s firing at $t_1$ still will occur; then, if we time-evolve forward according to the neuron laws, $B$ will fire at $t_2$, and $E$ will fire at $t_3$. So it is not the case that, had $C$'s firing at $t_1$ not occurred, then $E$'s firing at $t_3$ would not have occurred.

   $$\neg O(c) \not\Box\!\!\rightarrow \neg O(e)$$

   (where '$c$' is $C$'s firing at $t_1$ and '$e$' is $E$'s firing at $t_3$, and we've supposed that all that it takes for $E$'s firing at $t_3$ to occur is for $E$ to fire at $t_3$.)

   (b) Yet $C$'s firing at $t_1$ *caused* $E$'s firing at $t_3$.

   (c) So, causal dependence is not necessary for causation.

8. LEWIS (1973a)'s solution to the problem of preemption is to take causation to be, not causal dependence, but rather the *transitive closure* of causal dependence.

   LEWIS'S 1973 THEORY OF CAUSATION
   An event $c$ caused a distinct event $e$ iff either:

   (a) $e$ causally depends upon $c$;

   $$\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$$

   or

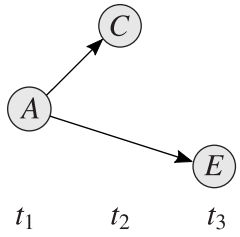
   (b) there is a sequence of events $d_1, d_2, \ldots, d_N$ such that $e$ causally depends upon $d_N$, $d_N$ causally depends upon $d_{N-1}, \ldots$, and $d_1$ causally depends upon $c$.

   $$\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(d_1) \wedge \neg O(d_1) \mathrel{\Box\!\!\rightarrow} \neg O(d_2) \wedge \cdots \wedge \neg O(d_N) \mathrel{\Box\!\!\rightarrow} \neg O(e)$$

9. In the case of preemption from figure 6.1, $E$'s firing at $t_2$ causally depends upon $D$'s firing at $t_2$; and $D$'s firing at $t_2$ causally depends upon $C$'s firing at $t_1$. Thus, according to LEWIS's account $C$'s firing at $t_1$ caused $E$'s firing at $t_2$.

10. LEWIS (1973a)'s theory is also able to deal with the cases of common causes we encountered earlier. Consider the neuron diagram in figure 6.2.

**Figure 6.2** A common cause




(a) Had $C$'s firing at $t_2$ not occurred, $A$'s firing at $t_1$ still would have (this is not a backtracker).

(b) So, had $C$'s firing at $t_2$ not occurred, $E$'s firing at $t_3$ still would have.

(c) And there is no sequence of events intermediate between $C$'s firing at $t_2$ and $E$'s firing at $t_3$ such that $E$'s firing causally depends upon the end of that sequence and the start of that sequence causally depends upon $C$'s firing.

(d) So $C$'s firing did not cause $E$'s firing, according to Lewis's account.

11. Lewis (1973a)'s account is able to deal with cases of common causes in part because it incorporates what's now generally called an *intervention* (or a 'Lewisian miracle').

### 6.2.1 Objections to Lewis's 1973 Theory

#### Probabilistic Causation

12. There are questions about how Lewis hopes to deal with probabilistic causation, since in those cases, the relevant counterfactuals will turn out either false or indeterminate, but these issues are dealt with in the 1986a 'postscripts'.

#### Late Preemption

13. The central problem that Lewis's account faces, as I see it, is that not all cases of preemption are like the one shown in figure 6.1. The case in figure 6.1 is a case of what Lewis (1986a) calls 'early preemption'. There are additionally cases of what he calls 'late preemption'.

(a) Compare the following two cases:

> Early Preemption
> Suzy wants to see mean old Mr. Wilson's window shattered. She'd rather not take the heat herself, so she goads Billy into throwing a rock at his window. If Billy wusses out and doesn't throw, then Suzy will throw her own rock and break the window. Billy throws, and the window shatters.

Late Preemption
Billy and Suzy both throw their rocks at exactly the same time, but Billy throws his a bit faster. Billy's rock hits the window, it shatters, and shortly thereafter Suzy's rock flies through the hole where the window used to be.

(b) In Early Preemption, while the window's shattering doesn't causally depend upon Billy's throw, there is an event which *does* causally depend upon Billy's throw—namely, his rock being in the air between Billy and the window—and the window's shattering causally depends upon that event, assuming—perhaps implausibly—that Suzy 'stands down' as soon as the rock leaves Billy's hand (getting cases of early preemption is harder than you might think at first). So there is a chain of causal dependence leading from Billy's throw to the window's shattering, and the 1973a account rules that Billy's throw caused the window to shatter.

(c) However, in Late Preemption, the window's shattering does not causally depend upon this intermediate event—the rock's being in midair between Billy and the window. For, if we imagine it away, Suzy's rock is still there to finish the job. And this is true all the way up until the time when the window shatters. So the 1973a account rules, incorrectly, that Billy's throw didn't cause the window to shatter.

14. One response to both Early and Late Preemption is to make the effect more *modally fragile*—to make it much easier for the effect of the window's shattering to fail to occur.

(a) Lewis (1986a) claims that this was his original thought about such cases. However, in 1986a, he believed that this strategy opens the door to too many spurious causes.

(b) Consider this alternative ending to the case:

Late Preemption (v2)
Suzy wants to see mean old Mr. Wilson's window shattered. She'd rather not take the heat herself, so she goads Billy into throwing a rock at his window. If Billy wusses out and doesn't throw, then Suzy will throw her own rock and break the window. Billy wusses out, Suzy throws, and the window shatters.

Here, then, we would have to say that Billy's *failure* to throw caused the window to shatter, since, had he thrown, the window would have shattered sooner or later than it actually did.

## Transitivity

15. Lewis's account commits him to the transitivity of causation (why?).

16. However, there soon emerged a large variety of counterexamples to the transitivity of causation. Here is a sampling:

Dog Bite
A (right-handed) terrorist plans to detonate a bomb inside a building on Monday. On Sunday, a dog bites their right hand. So, on Tuesday, they detonate the bomb with their left hand. The building explodes. (McDermott, 1995).

Boulder
A large boulder becomes dislodged, rolls down the hill, and careens towards a hiker. The boulder is large enough and fast enough that, if it hits the hiker, they will surely die. The hiker, seeing the boulder, ducks. The boulder flies over their head, and they survive unscathed. (attributed to an early draft of Hall 2004 by Hitchcock 2001b).

Switch
A train track branches off to the left and the right. Both tracks are headed for the station and both take an equal amount of time to get there. A switch controls whether a train will go off to the left or the right. You flip the switch, and a train is diverted off to the left at 12:00. The train arrives on time at 1:00. (Hall 2004)

(a) In Dog Bite, the dog biting the terrorist's right hand caused them to detonate the bomb with their left hand. Their detonating the bomb with their left hand caused the building to explode. However, the dog bite did not cause the building to explode.

(b) In Boulder, the boulder's becoming dislodged caused the hiker to duck. Their ducking caused them to survive. However, the boulder's becoming dislodged did not cause them to survive.

(c) In Switch, your flipping the switch to the left caused the train to be on the left track at 12:30. The trains being on the left track at 12:30 caused it to arrive at the station at 1:00, but your flipping the switch to the left did not cause the train to arrive at the station at 1:00.

17. Lewis remained unpersuaded by these examples to the end; you'll see his response when you read his 2004.

### 6.3 Lewis's 1986 Revision of the Counterfactual Theory

18. Firstly, Lewis wishes to deal with cases of probabilistic causation. He begins by introducing a new, probabilistic, account of causal dependence

Causal Dependence
An event *e causally depends upon* a distinct event *c*, which occurred just before $t_c$, iff:

$$\neg O(c) \boxdot\!\!\rightarrow Ch_{t_c}(O(e)) \ll Ch^{@}_{t_c}(O(e))$$

(where '$Ch^{@}$' is the actual world's chance function).

(a) That is: $e$ causally depends upon $c$ iff, had $c$ not occurred, the chance of $e$ would have been much lower than it actually is ('much lower', we are told, is to be understood in terms of 'lower by a large factor').

19. As before, causation is the ancestral, or the transitive closure, of causal dependence.

> Lewis's 1986 Theory of Causation (v1)
> An event $c$ caused a distinct event $e$ iff
> (a) $e$ causally depends upon $c$; or
> (b) There is a chain of events $d_1, d_2, \ldots, d_N$ such that $e$ causally depends upon $d_N, \ldots,$ and $d_1$ causally depends upon $c$.

20. Finally, to deal with cases of late preemption, Lewis (1986a) introduces the idea of *quasi-dependence*.

> Quasi-Dependence
> An event $e$ *quasi-depends* upon a distinct event $c$ iff the intrinsic character of the process leading from $c$ to $e$ is such that, in the great majority[3] of regions with this intrinsic character, either:
> (a) $e$ causally depends upon $c$; or
> (b) there is a chain of events $d_1, d_2, \ldots, d_N$ such that $e$ causally depends upon $d_N, \ldots,$ and $d_1$ causally depends upon $c$.

21. Then, the account is completed as follows:

> Lewis's 1986 Theory of Causation (v2)
> An event $c$ caused a distinct event $e$ iff
> (a) $e$ quasi-depends upon $c$; or
> (b) There is a chain of events $d_1, d_2, \ldots, d_N$ such that $e$ quasi-depends upon $d_N, \ldots,$ and $d_1$ quasi-depends upon $c$.

### 6.3.1 Problems with Probabilistic Causation

22. Lewis's 1986a theory of probabilistic causation, recall, re-defined causal dependence as counterfactual probability raising; and took causation to be the ancestral of causal dependence.

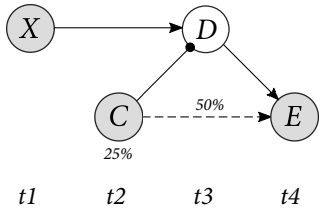
(a) That is, the account said that:

> Probabilistic Causation (Lewis 1986)
> An event $c$, which occurred just before $t_c$, caused a distinct event $e$, iff there is a chain of causal dependence leading from $c$ to $e$; where $d_2$ causally depends upon $d_1$ iff:

$$\neg O(d_1) \boxdot\!\!\rightarrow Ch_{t_c}(O(d_2)) \ll Ch^{@}_{t_c}(O(d_2))$$

[3] According to which measure? One given by "the variety of their surroundings"

**Figure 6.3** *C*'s firing caused *E*'s firing, though *E*'s firing does not causally depend upon *C*'s firing.




23. This account faces problems in both directions. The left-to-right direction fails in the neuron diagram in figure 6.3 (we've seen neuron diagrams like this before, when we talked about probabilistic theories of causation).

    (a) Here, *C*'s' firing at $t_2$ ($c$) caused *E*'s firing at $t_4$ ($e$).

    (b) However, the actual chance of *E*'s firing, at a time just after $t_2$, is

    $$Ch^{@}_{>t_2}(O(e)) = \frac{1}{2}$$

    (c) While, had *C* not fired at $t_2$, the chance of *E*'s firing at $t_4$ would have been 1.

    $$\neg O(c) \boxdot\!\!\rightarrow Ch_{t_2}(O(e)) = 1$$

    (d) So $e$ does not causally depend upon $c$.

    (e) Moreover, the only event intermediate between $c$ and $e$ is *D*'s failure to fire at $t_3$ ($d$). And, while $d$ does causally depend upon $c$, since

    $$\neg O(c) \boxdot\!\!\rightarrow Ch_{>t_2}(O(d)) = 0 \ll Ch^{@}(O(d)) = 1$$

    $e$ does not causally depend upon $d$, since, had *D* fired at $t_3$, *E* would have had a chance of 1 of firing at $t_4$.

    $$\neg O(d) \boxdot\!\!\rightarrow Ch_{>t_3}(O(e)) = 1 \not\ll Ch^{@}_{>t_3}(O(e)) = 1/2$$

    (f) So there is no chain of causal dependence leading from *C*'s firing to *E*'s firing, and therefore, Lewis's 1986a account will rule, incorrectly, that *C*'s firing did not cause *E*'s firing.

24. The right-to-left direction fails for reasons familiar from our discussion of probabilistic causation. Not all probability raisers (not even all *counterfactual* probability raisers) are causes (*cf.* Lewis's discussion of the bombs on the two planes in his 2004).

    (a) Consider the neuron diagram shown in figure 6.4.

**Figure 6.4** *C*'s firing did not cause *E*'s firing, though *E*'s firing does causally depend upon *C*'s firing.

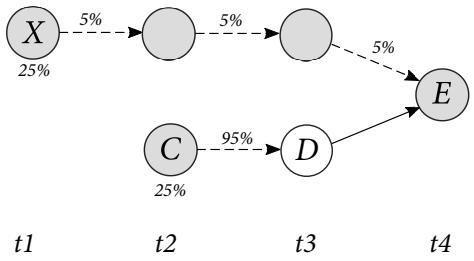
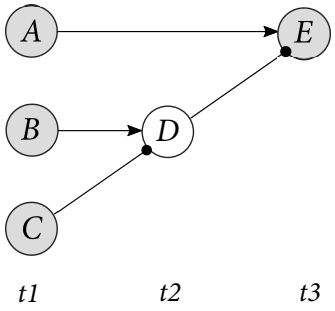



**Figure 6.5** A case of double prevention.




(b) *E*'s firing at $t_4$ ($e$) causally depends upon *C*'s firing at $t_2$ ($c$), since, had $C$ not fired at $t_2$, the chance just after $t_2$ of *E*'s firing at $t_4$ would have been substantially lower than it in fact was.

$$\neg O(c) \boxdot\!\!\rightarrow Ch_{>t_2}(O(e)) = 0.000125 \ll Ch^{@}_{>t_2}(O(e)) = 0.95$$

(c) Nevertheless, *C*'s firing at $t_2$ did not cause *E*'s firing at $t_4$.

### 6.3.2 Problems with Quasi-Dependence

25. One problem Lewis sees with his 1986a quasi-dependence account is that it was based on the idea that causation is an *intrinsic* relation. However, if you accept that counterfactual dependence is sufficient for causation, then you must deny that causation is an intrinsic relation.

(a) Consider a case of *double prevention* (figure 6.5)

(b) The counterfactual theorist is committed to *C*'s firing at $t_1$ causing *E*'s firing at $t_3$; for, had $C$ not fired at $t_1$, $E$ would not have fired at $t_2$.

(c) However, had $B$ not fired, then $C$'s firing at $t_1$ would *not* have caused $E$'s firing at $t_3$.

(d) But whether $B$ fires at $t_1$ is an extrinsic fact.

(e) So, whether $C$'s' firing at $t_1$ caused $E$'s firing at $t_3$ is an extrinsic fact.

### 6.3.3 Trumping Preemption

26. A further problem Lewis cites for both quasi-dependence and his 1973a accounts of causation are cases of so-called *trumping preemption*. Here are three examples:

> Major/Sergeant
> The Major outranks the Sergeant, and the soldiers always follow the order of the highest-ranking officer. Both the Major and the Sergeant order the soldiers to advance. They advance.
>
> Merlin/Morgana
> The laws of magic say that the first spell cast on any given day will take effect at midnight. In the morning, Merlin casts a spell to turn the prince into a frog. In the evening, Morgana casts a spell to turn the prince into a frog. At midnight, the prince turns into a frog.
>
> White Field/Black Field
> There are two kinds of fields: black and white fields, which act on particles according to the following law: particles are accelerated in the direction of the *darkest* field which acts upon them. (See figure 6.6.) So, if a particle goes through a white field with a source on the left, it will accelerate towards towards the left (6.6(a)); if it goes through a white field with a source on the right, it will accelerate towards the right (6.6(b)); if it goes through a white and a black field, it will accelerate in the direction of the source of the black field, no matter where the source of the white field is (6.6(c) and 6.6(d)).

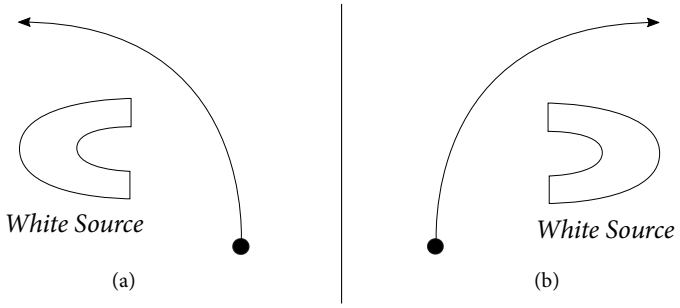
(a) In Major/Sergeant, it was the *Major*'s order which caused the men to march, and not the Sergeant's. However, the advance counterfactually depends upon neither order.

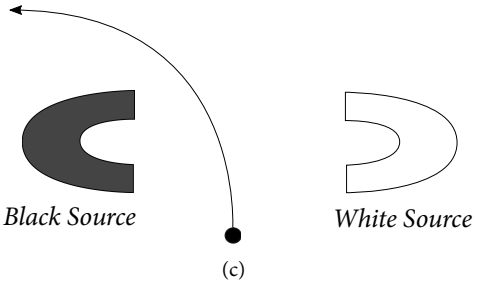
(b) In Merlin/Morgana, it was *Merlin's* spell that caused the prince to turn into a frog, and not Morgana's. However, the prince's turning into a frog counterfactually depends upon neither spell.

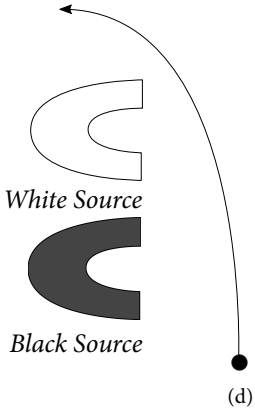
(c) In figure (6.6(d)), it was the *black source* which caused the particle to accelerate to the left, and not the white source. However, the particle's acceleration counterfactually depends upon neither source.

27. What's meant to be distinctive about *trumping preemption*, as opposed to ordinary cases of preemption, is that, in ordinary cases of preemption, there is thought to be a *cut*, or a *break*, in the causal process leading from the preempted backup to the effect—unlike in cases of symmetric overdetermination, the process leading from

**Figure 6.6** Black Fields and White Fields



the preempted backup does not run through to completion. Many counterfactual theorists attempted to use this break in the causal process to distinguish the cause from the preempted backup. However, in cases of trumping preemption, the causal process leading from the trumped potential cause (the Sergeant's orders, Morgana's spell, and the white source) run to completion.

(a) In particular, for Lewis's original 1973a account, there is no intermediate event on which the effect depends and which depends in turn upon the trumped potential cause.

(b) And, in the case of Lewis's revised 1986a *quasi-dependence* account, there are intrinsic duplicates of the causal process leading from the trumped potential cause to the effect in which that effect causally depends upon the trumped potential cause. So the *quasi-dependence* account rules, incorrectly, that the trumped potential causes are genuine causes.

### 6.4 Lewis's 2000 Influence Account

28. Another objection to the 1973a and 1986a accounts is that it requires us to draw a sharp line between those possibilities in which an event occurs at a different time, or in a different manner, and those possibilities in which the event *fails to occur*, and another event takes its place.

    (a) Lewis expresses skepticism about the possibility of drawing this line in general. Our linguistic practice simply does not draw such sharp lines.

    (b) So, Lewis decides to shift his focus from the occurrence and non-occurrence of events to all the various *alterations* in the time and manner of an event (whether those alterations are alterations in which the event fails to occur, or alterations in which the event occurs, but at a different time or in a different manner).

        i. Note that, as Lewis is using the term, every event counts as an alteration of itself (the unaltered alteration).

29. Using this notion of an *alteration*, Lewis gives an account of *causal influence*:

    > Causal Influence
    > An event $c$ *influences* a distinct event $e$ iff there is a substantial range of not-too-distant alterations of $c$, $c_1, c_2, \ldots, c_N$, and a range of alterations of $e$, $e_1, e_2, \ldots, e_N$ (at least some of which differ) such the $c_i$'s counterfactually pattern with the $e_i$'s—that is, such that
    >
    > $$O(c_1) \boxdot\!\!\rightarrow O(e_1) \wedge O(c_2) \boxdot\!\!\rightarrow O(e_2) \wedge \cdots \wedge O(c_N) \boxdot\!\!\rightarrow O(e_N)$$

    (a) Note that, while Lewis requires that *some* of the $e_i$'s differ, he does not require that *all* of them do.

30. Does causal dependence entail causal influence? It's not clear, but I believe the answer is 'yes'.

(a) One thing that makes the question hard is that Lewis does not specify what makes a range of alterations of $c$ "substantial". However, in many of his examples (for instance, the 'inert' neuron diagram on page 98), he appears content to list just a single alteration of the cause (one in which the cause does not occur) with which two alterations of the effect counterfactually pattern, and straightaway conclude that the cause influences the effect.

(b) So let's assume that, if $\overline{c}$ is the alteration of $c$ which *would* occur, were $c$ to not occur, then $c$ and $\overline{c}$ constitutes a substantial range of not-too-distant alterations of $c$.

(c) Then, since $O(\overline{c}) \mathrel{\Box\!\!\rightarrow} \neg O(c)$ and $\neg O(c) \mathrel{\Box\!\!\rightarrow} O(\overline{c})$, it follows on Lewis (1973b)'s semantics for the counterfactual that

$$\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e) \quad \Rightarrow \quad O(\overline{c}) \mathrel{\Box\!\!\rightarrow} O(\overline{e})$$

And since

$$O(c) \mathrel{\Box\!\!\rightarrow} O(e)$$

follows from strong centering, we have that

$$O(c) \mathrel{\Box\!\!\rightarrow} O(e) \wedge O(\overline{c}) \mathrel{\Box\!\!\rightarrow} O(\overline{e})$$

(d) So causal dependence entails causal influence (though the converse is false).

31. Lewis (2000) says that causation is the ancestral of influence.

> Causation (Lewis (2000))
> An event $c$ caused a distinct event $e$ iff either $c$ influenced $e$ or there is a sequence of events $d_1, d_2, \ldots, d_N$ such that:
> (a) $c$ influenced $d_1$;
> (b) $d_1$ influenced $d_2$;
> $\vdots$
> ($n$) $d_{N-1}$ influenced $d_N$; and
> ($m$) $d_N$ influenced $e$.

32. Lewis believes that this account allows him to get the right verdict both in cases of late preemption and cases of trumping preemption.

(a) Consider Late Preemption—Suzy and Billy both throw, Suzy's rock hits, and the window shatters before Billy's rock arrives. While the window's shattering does not depend upon Suzy's throw, alterations of the window's shattering do counterfactually pattern with alterations of Suzy's throw. So Suzy's throw caused the window to shatter, according to the influence account. Not so (or near enough) for Billy's throw.

(b) Consider Major/Sergeant. While alterations of the Major's order in which he gives *no* order do not counterfactually pattern with alterations of the soldier's advancing, alterations of the Major's order in which he gives a *different* order *do* counterfactually pattern with alterations of the soldier's advancing. Not so (or near enough) for the Sergeant's order

33. Because causal dependence entails causal influence (but not vice versa), the influence account lets in *more* causes than does the 1973a dependence account. This is good, since the problems with the 1973a account were that it didn't let in enough causes.

34. However, there's a worry that the 2000 influence account lets in *far too many* causes.

    (a) Billy and his rock exert minor gravitational and electromagnetic forces on the window, and do make (undetectable, perhaps) differences to the manner in which the window shatters.

    (b) Thus, Billy's throw does influence the window's shattering after all.

35. Lewis (2004)'s response: "Well—these differences made by spurious causes are negligible, so surely we are entitled to neglect them?" (p.89)

    (a) The idea is this: the degree to which $c$ is appropriately cited as a cause of $e$ in a causal claim is proportional to its relative degree of influence on $e$.

    (b) That is: if $c_1$ has *more* of an influence on $e$ than $c_2$, then it will be more appropriate to cite $c_1$ as a cause of $e$ in a causal claim than it will be to cite $c_2$ as a cause of $e$ in a causal claim.

36. Objections to this method of dealing with these spurious causes emerged in Schaffer (2001) and Strevens (2003). Here's a similar case from Gallow (2015):

    > Poison/Anesthesia
    > Sabeen tells you truthfully that she plans to slip a fatal poison into Stephanie's drink. You are unable to warn Stephanie, and you do not 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.

    (a) In this case, the poison had a relatively small influence on Stephanie's death. 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.)

    (b) And the anesthetic had a relatively large influence on Stephanie's death. 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.

    (c) Still, it seems very inappropriate to say

i. # Your pouring the anesthetic caused Stephanie to die.

and appropriate to say

ii. Sabeen's pouring the poison caused Stephanie to die.

(d) So: Lewis's strategy for ruling out spurious causes does not rule out all of the spurious causes. It leaves behind some, like, *e.g.*, your pouring of the anesthetic.

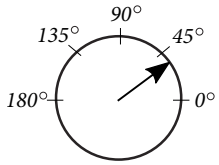
# 7 De Facto Dependence and Counterfactual Counterfactual Theories

## 7.1 Structural Equations Models

1. A *structural equations model* (or, alternatively, a *causal model*) $\mathcal{M}$ is a 4-tuple $\langle \mathcal{U}, \mathbf{u}, \mathcal{V}, \mathcal{E} \rangle$ of:[1]
   (a) A vector $\mathcal{U} = [U_1, U_2, \ldots, U_M]'$ of *exogenous* variables;
   (b) A *context* $\mathbf{u} = [u_1, u_2, \ldots, u_M]$, which is just a vector which assigns a value to each exogenous variable in $\mathcal{U}$;
   (c) A vector $\mathcal{V} = [V_1, V_2, \ldots, V_N]'$ of *endogenous* variables; and
   (d) A vector $\mathcal{E} = [\phi_{V_1}, \phi_{V_2}, \ldots, \phi_{V_N}]'$ of *structural equations*—one for each endogenous $V_i \in \mathcal{V}$.
      i. A vector is just an ordered list of items. (It's helpful for the variables to have an order because we want to talk about assignments of values to all of the exogenous/endogenous variables, and we'll want to be able to clearly associate which value is associated with which variable.)
      ii. Note that my presentation here is non-standard in the following respect: contexts are not usually specified as being a part of the structural equations model.
2. For our purposes, let's understand a *variable* in the following way: a variable is a (perhaps partial) function from the set of possibilities $\mathscr{W}$ to real numbers $\mathbb{R}$.
   (a) Thus, we might have a variable $A$ for the *angle* (in degrees) of the volume knob on the radio. This variable maps those possibilities at which the radio exists to real numbers between 0 and 180. It maps a possibility to $x$ iff, at that possibility, the angle of the volume knob on the radio is $x$ (at the relevant time).
   (b) We might also have a variable $D$ for the volume (in *decibels*) of the sound coming out of the radio's speaker. This variable maps those possibilities at which the radio exists to real numbers. It maps a possibility to $x$ iff, at that possibility, the decibel level of sound coming out of the radio's speaker is $x$.

[1] Notation: I'll use caligraphic and boldface letters for vectors of variables, capital letters for variables, lowercase letters for the values of those variables, and '$\phi_V$' for the structural equation associated with the endogenous variable $V$.

**Figure 7.1** The angle of the volume knob on the radio.




(c) We can write '$V_w = v$' to mean that the value of the function $V$, given the argument $w$, is $v$. Or, alternatively, that the *value* of the variable $V$, *at* $w$, is $v$.

i. Thus, $A_w = 42.4$ tells us that, at $w$, the angle of the volume knob (at the relevant time) is 42.4°.

ii. And $D_w = 137.6$ tells us that, at $w$, the decibel level of the sound coming out of the speaker (at the relevant time) is 137.6 decibels.

3. *Structural equations* provide us with equations relating the values of variables. For instance, it could be that the decibel level of the sound coming out of the speaker ranges from 0 (if the knob is at 180°) to 180 (if the knob is at 0°). Then, the following *structural equation* may be true:

$$D := 180 - A$$

(a) The equation is *structural* because it tells us more than just that the decibel level of sound coming out of the radio is a function $180 - A$ of the angle of the volume knob. It additionally tells us that the value of $D$ is *determined by* the value of $A$ (and not *vice versa*).

(b) The equation '$D = 180 - A$' could be re-written as '$A = 180 - D$' or '$A + D = 180$'. The *structural* equation '$D := 180 - A$' *cannot* be re-written as '$A := 180 - D$'. It matters which variable is to the left of the ':=' sign. That's why I've used ':=', rather than '='.[2]

4. The following, then, is a simple structural equations model:

$$\mathcal{M} = \left\langle \begin{array}{rl} \mathcal{U} & = [A] \\ \mathcal{V} & = [D] \\ \mathcal{E} & = [D := 180 - A] \\ \mathbf{u} & = [42.4] \end{array} \right\rangle$$

(a) This model tells us that the decibel level of the sound is determined by the angle of the volume knob according to the equation $D := 180 - A$, and that, actually, the volume knob is at an angle of 42.4°.

5. We can represent some of what the structural equations model tells us with the aid of a *causal graph*.

[2] I think the first use of this notation was in Weslake (forthcoming). Other authors will use '$c =$' or '$\Leftarrow$'.

(a) A causal graph tells us which variables have their values immediately determined by which others. For every structural equation, we draw arrows with their heads at the variable on the left-hand-side of the structural equation and with their bases at the variables on the right-hand-side of the structural equation.

(b) For instance, we can represent some of the information in our model above with the following graph:

$$A \longrightarrow D$$

(c) This graph tells us that $A$ is an exogenous variable (since it has no arrows leading into it), $D$ is an endogenous variable (since it does have arrows leading into it), and that $D$'s value is determined by $A$'s value—though it doesn't tell us precisely *how*. It is consistent with all that this causal graph tells us that $D$'s value is determined by $A$'s value according to

$$D := A$$

or

$$D := 0.2(180 - A)$$

(d) Notice that, since the model gives us the value of $A$, it allows us to solve for the value of $D$. Since $A = 42.4$, $D = 180 - 42.4 = 137.6$.

(e) This is true in general, so long as $\mathcal{M}$ is *acyclic*.

**Acyclicity.** *A structural equations model $\mathcal{M}$ is* ACYCLIC *iff, in the causal graph of $\mathcal{M}$, there are no directed paths leading from a variable to itself—that is, in the causal graph, there is no variable $V$ and a sequence of arrows directed tail-to-tip such that $V \to X_1 \to X_2 \to \cdots \to X_N \to V$.*

(f) If $\mathcal{E}$ is *acyclic*, then a structural equations model $\mathcal{M}$ will entail the values of all of the endogenous variables.

(g) If '$\phi$' is a proposition true of the model $\mathcal{M}$, then we will write:

$$\mathcal{M} \models \phi$$

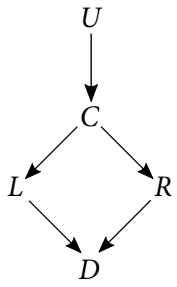
(h) Thus, in the model above,

$$\mathcal{M} \models D = 137.6$$

6. A causal graph can involve many more variables and equations than this. For a slightly more complicated instance, (PEARL, 2000, p. 207) gives the following example:

**Firing Squad.** *There is a two-man firing squad; and L, R, C, D, and U are variables which take the value* 1 *if their corresponding propositions are true and take the value* 0 *if their corresponding propositions are false. The corresponding propositions are:*

**Figure 7.2** Causal graph for the structural equations model from **Example 1**.




$U$ – *the court orders the execution*
$C$ – *the captain gives a signal*
$L$ – *the left rifleman shoots*
$R$ – *the right rifleman shoots*
$D$ – *the prisoner dies*

*$U$ is exogenous; and all other variables are endogenous. The relevant equations are the following:*

$$\mathcal{E} = \begin{bmatrix} C & := U \\ L & := C \\ R & := C \\ D & := L \vee R \end{bmatrix}$$

*(where '$\circ \vee *$' is the truth-function* $\min\{\circ, *\}$*.) In fact, the court orders the execution.*

The causal model from **Example 1** has the causal graph shown in figure 7.2.

(a) In this model, the context is $\mathbf{u} = [1]$, which assigns the value 1 to the (only) exogenous variable $U$. Using the equations in $\mathcal{E}$, we can then solve for every other variable in the model. Doing so tells us that

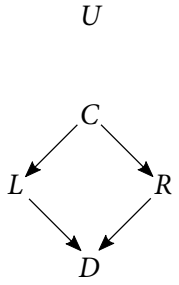
$$\mathcal{M} \models C = L = R = D = 1$$

7. Causal models additionally provide a semantics for non-backtracking counterfactual conditionals. The idea behind this semantics is essentially the idea behind Lewis (1979)'s idea of a tiny miracle. If we want to evaluate the counterfactual 'had the captain not given a signal, the prisoner would not have died', in causal model $\mathcal{M}$, then we *mutilate* $\mathcal{M}$ by *removing* $C$'s structural equation, thus rendering $C$ *exogenous*; and we update the context so that it assigns the value 0 to the (now exogenous) variable $C$.

(a) Call this mutiliated model '$\mathcal{M}_{C=0}$'. We then have that

$$\mathcal{M}_{C=0} = \langle \mathcal{U}_{C=0}, \mathcal{V}_{C=1}, \mathcal{E}_{C=0}, \mathbf{u}_{C=0} \rangle,$$

**Figure 7.3** Causal graph for the 'mutilated' model $\mathcal{M}_{C=0}$.




where $\mathcal{U}_{C=0} = [U, C]$, $\mathbf{u}_{C=0} = [1, 0]$, $\mathcal{V}_{C=0} = [L, R, D]$, and

$$\mathcal{E}_{C=0} = \begin{bmatrix} L & := C \\ R & := C \\ D & := L \vee R \end{bmatrix}$$

The causal graph associated with the mutilated model $\mathcal{M}_{C=0}$ is shown in figure 7.3.

(b) Now, check whether '$D = 0$' is true in this new causal model $\mathcal{M}_{C=0}$. When we run through the equations, we see that it is—that is, we see that:

$$\mathcal{M}_{C=0} \models D = 0$$

(c) This tells us that, *in the original model*, the counterfactual 'had the captain not given a signal, the prisoner would not have died' is true.

$$\mathcal{M} \models C = 0 \mathrel{\square\!\!\rightarrow} D = 0$$

8. This same procedure is used to evaluate counterfactual conditionals in general.

(a) Suppose that we have a vector of variables $\mathbf{X}$ (all of which appear in our model $\mathcal{M}$), and an assignment of values to those variables, $\mathbf{x}$, and we wish to know whether

$$\mathcal{M} \models \mathbf{X} = \mathbf{x} \mathrel{\square\!\!\rightarrow} \phi$$

(where $\phi$ is some arbitrary proposition about the values of variables in our model.)

(b) We get the multilated model $\mathcal{M}_{\mathbf{X}=\mathbf{x}}$ by:

i. adding all of the endogenous variables in $\mathbf{X}$ to $\mathcal{U}$;
ii. adding the assignment $\mathbf{x}$ to $\mathbf{X}$ to the context $\mathbf{u}$;
iii. removing the endogenous variables in $\mathbf{X}$ from $\mathcal{V}$; and
iv. removing the structural equations of the endogenous variables in $\mathbf{X}$ from $\mathcal{E}$.

Relying on your intuitive understanding of what I mean by '+' and '−', we can think about the effect of mutilating $\mathcal{M}$ with the following symbolic instructions:

$$\begin{aligned}\mathcal{U}_{\mathbf{X}=\mathbf{x}} &= \mathcal{U} + \mathbf{X} \\ \mathbf{u}_{\mathbf{X}=\mathbf{x}} &= \mathbf{u} + \mathbf{x} \\ \mathcal{V}_{\mathbf{X}=\mathbf{x}} &= \mathcal{V} - \mathbf{X} \\ \mathcal{E}_{\mathbf{X}=\mathbf{x}} &= \mathcal{E} - \phi_{\mathbf{X}}\end{aligned}$$

(c) Then, our general semantics for counterfactual conditionals tells us that

$$\mathcal{M} \models \mathbf{X} = \mathbf{x} \boxdot\!\!\rightarrow \phi \iff \mathcal{M}_{\mathbf{X}=\mathbf{x}} \models \phi \qquad (\boxdot\!\!\rightarrow)$$

9. This semantics for counterfactual conditionals is not equivalent to the Lewis/Stalnaker semantics. One of the more interesting departures was pointed out by Briggs (2012): given this semantics, counterfactual conditionals do not satisfy *modus ponens* (the rule of inference which says that, from $\phi$ and $\phi \boxdot\!\!\rightarrow \psi$, you may infer $\psi$).

   (a) The counterexample from Briggs uses the causal model from **Example 1**. She notes that, in that model, both of the following claims are true:

$$\begin{aligned}&L = 1 \boxdot\!\!\rightarrow (C = 0 \boxdot\!\!\rightarrow D = 1) \\ &L = 1\end{aligned}$$

   However, it is false that

$$C = 0 \boxdot\!\!\rightarrow D = 1$$

   i. *Cf.* the counterexamples to modus ponens from McGee (1985).

10. Note that the semantics for counterfactual conditionals only tells us what to do if our antecedents specify assignments of values to variables. It is silent on what we should do about, *e.g.*, antecedents which themselves contain counterfactuals, or antecedents which contain *disjunctions* of variable values, like, e.g., $X = x \vee Y = y$. Briggs (2012) and Huber (2013) have suggestions, but we won't be concerned with them here.

11. Our focus here will be on attempts to utilize structural equations models to provide an account of singular causation.[3]

   (a) Let's call any account of singular causation formulated in terms of structural equations model a 'structural' theory of causation.

12. Structural theories of causation incur two theoretical burdens:

[3] In the literature, this is increasingly known as 'actual causation', but I hate the name, so I'll stick to 'singular' and 'token' here.

(a) First, they must say something about what it takes for a particular structural equations model to be *correct* at a world of evaluation, $w$.[4] Call this *a theory of structural determination*.

(b) Secondly, they must say something about when what it takes for a variable value $C = c$ to cause a distinct variable value $E = e$ *within* a given structural equations model.[5] Call this *a theory of singular causation*.

13. We will mostly be ignoring some of the complexities that arise when you try to give a theory of structural determination. However, some of the comments from Hitchcock (2001b) will provide us with enough guidance to think about whether a particular structural equations model is appropriate for a given situation. Hitchcock tells us that a structural equation $V := \phi_V(W, X, \ldots, Z)$

> ...encodes a set of counterfactuals of the following form:
>
> > If it were the case that $W = w, X = x, \ldots,$ and $Z = z$, then it would be the case that $V = \phi_V(w, x, \ldots, z)$.
>
> ..Each equation in $\mathcal{E}$ encodes counterfactual information. Note, however, that the equations $\mathcal{E}$ do not directly represent *all* counterfactuals that are true of the system. Rather, $\mathcal{E}$ is a set of *fundamental* equations from which all other counterfactuals may be derived. ...A system of structural equations is an elegant means for representing a whole family of counterfactuals of just the sort that Lewis's counterfactual theory of causation depends upon. The correctness of a set of structural equations, and of the corresponding graph, depends upon the truth of these counterfactuals.[6]

(a) That is: a causal model will entail a whole slew of (non-backtracking) counterfactuals. If all of these counterfactuals are true, then the causal model is correct.[7]

## 7.2 De Facto Dependence

14. To see what extra theoretical toe-holds the formalism of structural equations buys us, let's think back to cases of *early preemption*. (See figure 7.4).

15. Yablo (2002, 2004) noticed the following about cases like these: while $E$'s firing does not counterfactually depend upon $C$'s firing. Had $C$ not fired, $D$ would have, and $E$ would have fired all the same. But, as it happens, $D$ *didn't* fire. And, if we hold fixed the true fact that $D$ didn't fire, then $E$'s firing *does* counterfactually depend upon $C$'s firing.

---

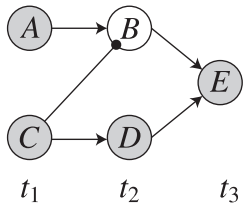
[4] See Hall (2007) and Paul & Hall (2013) for criticisms of structural theories of causation's failure to provide such an account. See Halpern & Hitchcock (2010) and Gallow (2016) for discussion and proposals.

[5] Many accounts actually allow conjunctions of variable values to be causes, and arbitrary Boolean combinations of variable values to be effects; however, we won't worry about such complications here.

[6] Hitchcock (2001b, p. 280, p. 283, & p. 284, with notational changes).

[7] See Gallow (2016) for arguments against this account of structural determination.

**Figure 7.4** Early Preemption




(a) If you can find some true fact of the appropriate sort, $G$ (which Yablo calls the 'ground' of the causal relation) such that

$$(\neg O(c) \wedge G) \; \square\!\!\rightarrow \neg O(e)$$

then Yablo says that *e de facto depends* upon *c*. And Yablo proposes that causation be analyzed in terms of *de facto dependence*.

(b) The requirement that the true fact $G$ be of the appropriate sort is an important restriction. Without it, we would be able to conclude that every event caused every other event. For, if $d$ and $e$ are any two events which actually occur, then $O(d) \leftrightarrow O(e)$ will be true. And, holding this true fact fixed, had $d$ not occurred, $e$ would not have occurred either,

$$(\neg O(d) \wedge (O(d) \leftrightarrow O(e))) \; \square\!\!\rightarrow \neg O(e)$$

This would be so even if $d$ were the event of Michelle Bachmann praying for Trump's victory and $e$ were the event of Trump getting elected. But Bachmann's prayers did not cause Trump to win (or, at least, the question is not so easily settled). So we need to place some restrictions of what kinds of grounds $G$ are *suitable*.

(c) Yablo does not provide an account of which grounds are suitable—it's not for nothing that his article was titled "Advertisement for a Sketch of an Outline of a Prototheory of Causation"; but the idea is promising.

16. We can use structural equations models to implement Yablo's proposal. The basic idea, which we can call 'De Facto', is this:

> De Facto
> $C = c$ caused $E = e$ in causal model $\mathcal{M}$ iff,
>
> $$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models C \neq c \; \square\!\!\rightarrow E \neq e$$
>
> for suitable variables $\mathbf{G}$,

(a) A word on notation: we're using '$\mathbf{G}_@$' to represent the actual values of the variables in $\mathbf{G}$—that is, the values that $\mathbf{G}$ takes on in the actual context, $\mathbf{u}$.

(b) The antecedent and the consequent in De Facto are propositions of the form '$V \neq v$'. The structural equations semantics for the counterfactual ($\Box\!\!\rightarrow$) does not tell us how to evaluate antecedents of that form; so we'll have to say something more about how to interpret this counterfactual.

i. First, note that, if $C$ and $E$ are binary variables standing for the occurrence or non-occurrence of events (and both of those events actually occurred—that is, $C = E = 1$), then the counterfactual in De Facto is to be understood as saying that $C = 0 \Box\!\!\rightarrow E = 0$.

ii. However, if $C$ and $E$ have 3 or more values, then we have choices. We could say that the counterfactual in De Facto is true iff there is *some* value of $C$ other than $c$ for which the counterfactual holds; that is,

$$C \neq c \Box\!\!\rightarrow E \neq e \iff \exists c' \neq c(C = c' \Box\!\!\rightarrow E \neq e)$$

iii. Alternatively, we could think that certain values of $C$ are *more similar* to $c$ than other values of $C$ (or that there are certain *default* values of $c$), and that those are the correct ones to be using to evaluate the counterfactual in De Facto.

iv. Finally, we might embrace a form of *contrastivism* about causation, according to which the causal relation is 4-place. On this view, causal claims are fundamentally of the form "$C = c$, rather than $C = c'$, caused $E = e$, rather that $E = e'$ (we'll talk more about this view later on). If we are contrastivists, then we'll want to re-write De Facto as follows:

> De Facto (Contrastive)
> According to the causal model $\mathcal{M}$, $C = c$, rather than $C = c'$, caused $E = e$, rather than $E = e'$, iff, for some suitable $\mathbf{G} \subseteq \mathcal{U} \cup \mathcal{V}$,
> $$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models C = c' \Box\!\!\rightarrow E = e'$$

v. For the examples we'll be considering here, all variables will be binary; so all of the alternatives above will be equivalent.

(c) Secondly, note that De Facto only defines causation *relative to* a particular causal model. What if there are multiple correct causal models such that, when we apply our definition of causation to those two models, they yield two different verdicts about whether $C = c$ caused $E = e$? Here, there are a variety of options. We could say:

i. There could never be such a pair of models.[8]

ii. The claim '$C = c$ caused $E = e$' is true iff there is *at least one* correct model $\mathcal{M}$ such that, according to $\mathcal{M}$, $C = c$ caused $E = e$.[9]

iii. The claim '$C = c$ caused $E = e$' is really just a claim of the form '$C = c$ caused $E = e$ in model $\mathcal{M}$'. When models are not explicitly mentioned (as

[8] I believe that Halpern (ms) takes this line.

[9] See Blanchard & Schaffer (forthcoming) for further discussion of this option.

they almost never are), context will settle which model we are implicitly talking about.[10]

iv. The claim '$C = c$ caused $E = e$' is to be evaluated, not just relative to a world, but additionally to a particular *causal model*. Thus, the same claim could be true relative to one causal model and false relative to another.

17. First, let's consider which variables are suitable for inclusion in $\mathbf{G}$.

18. A first-pass account is to suggest that *any* vector of variables $\mathbf{G}$ is suitable. This is equivalent to the suggestion made by Hitchcock (2001b). Call this theory 'Hitchcock (v1)'.

    Hitchcock (v1)
    According to the causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some vector of variables $\mathbf{G}$ such that

    $$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models C \neq c \; \square\!\!\rightarrow E \neq e$$

    (a) Note that we don't have to worry about trivial de facto dependence, since, if we let $\mathbf{G}$ contain only $C$, then we get that

    $$\mathcal{M}_{C=c} \models C \neq c \; \square\!\!\rightarrow E \neq e$$

    which will hold when and only when $\mathcal{M} \models C \neq c \; \square\!\!\rightarrow E \neq e$. Since counterfactual theorists accept that counterfactual dependence is sufficient for causation, this is no trouble.

    (b) Similarly, if $\mathbf{G}$ contains only $E$, then we get that

    $$\mathcal{M}_{E=e} \models C \neq c \; \square\!\!\rightarrow E \neq e$$

    which will never hold.

    (c) If $\mathbf{G}$ contains a variable blocking every path from $C$ to $E$, then it will definitely be false that

    $$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models C \neq c \; \square\!\!\rightarrow E \neq e$$

    (It is for this reason that Hitchcock (v1) is equivalent to Hitchcock (2001b)'s formulation in terms of an *active path*, and holding the *off path* variables fixed.)

19. Hitchcock (v1) is able to handle our case of early preemption (see figure 7.4). To see this, let's first construct a causal model of the neuron diagram in figure 7.4. This is shown in figure 7.5 (there, each variable is binary, and it takes the value 1 if the associated neuron fires at its designated time and takes the value 0 otherwise).

    (a) For this model, $\mathcal{M}$, it is true that

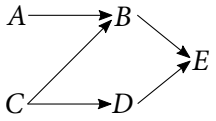
    $$\mathcal{M}_{B=0} \models C = 0 \; \square\!\!\rightarrow E = 0$$

    So, according to BDFA, $C$'s firing caused $E$'s firing. So we get this case right.

[10] See Menzies (2004). Additionally, Halpern & Pearl (2005) and Halpern & Hitchcock (2010) appear to endorse this option (though they may also be going in for a form of relativism—it's not clear from what they say).

**Figure 7.5** A structural equations model of the case of Early Preemption from figure 7.4. (The context is $A = C = 1$.)

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




20. Similarly, there are two ways (discussed in Halpern & Pearl (2005, Example 4.2)) for treating the case of Late Preemption. Consider first the structural equations model in figure 7.6.

    (a) In this model, $\mathcal{M}$, it is true that

    $$\mathcal{M}_{BH=0} \models ST = 0 \boxdot\!\!\rightarrow W = 0$$

    So whether the window shatters de facto depends upon whether Suzy throws.

21. Alternatively, we could model the scenario of Late Preemption using just the variables $BT$, $ST$, and $W$. However, we will then have to understand these variables as being time-indexed, and we'll let $t_0$ be the time Suzy threw, $t_1$ be the time Billy threw, and let $W_{t_n}$ stand for the proposition "the window is shattered at $t_n$". We could then model things with the system of equations shown in figure 7.7. In that model, $\mathcal{M}$, it is true (without even the need for de facto dependence) that

    $$\mathcal{M} \models ST_{t_0} = 0 \boxdot\!\!\rightarrow W_{t_1} = 0$$

    So Suzy caused the window to shatter at $t_1$. Moreover, Billy's throw did not cause the window to shatter (at any time), since $W_{t_2}$ does not de facto depend upon $B_{t_1}$.

22. The case which is *very* often cited as a reason to abandon the de facto dependence approach of Hitchcock (v1)[11] is a case of symmetric overdetermination. For instance, consider the case from figure 7.8. (There, $A$, $B$, and $F$ are binary variables with the natural interpretation—$A = 1$ if the first arsonist throws a match and is 0 if they don't; $B = 1$ if the second arsonist throws a match and is 0 if they don't; and $F = 1$ if the factory is set on fire and 0 if it is not.)

    (a) In this scenario, $F = 1$ does not de facto depend upon either $A = 1$ or $B = 1$. So Hitchcock (v1) tells us that neither $A = 1$ nor $B = 1$ caused $F = 1$.

    (b) We've seen cases with this structure before. Both Mackie and Lewis were happy to say that neither $A$ nor $B$ were individually causes of the factory catching on fire. Mackie wanted to insist that, nevertheless, their *disjunction* was still a cause of the fire, and Lewis wanted their *fusion* to be a cause of the fire.

[11] See, just for a sample, Hitchcock (2001b), Halpern & Pearl (2005), Halpern & Hitchcock (2010, forthcoming), and Weslake (forthcoming).

**Figure 7.6** Late Preemption

Both Suzy and Billy throw their rocks at the window at the same time. Suzy stands a bit closer, so her rock hits the window first, and the window shatters. Billy's rock flies through the space where the window used to be. The following binary variables stand for the truth-values of their associated propositions:

$ST$ – Suzy throws her rock
$BT$ – Billy throws his rock
$SH$ – Suzy's rock hits the window
$BH$ – Billy's rock hits the window
$W$ – The window shatters

$$SH := ST$$
$$BH := BT \wedge \neg SH$$
$$W := SH \vee BH$$

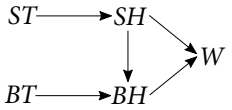



**Figure 7.7** Late Preemption with time-indexed variables

$$W_{t_1} := BT_{t_0} \vee ST_{t_0}$$
$$W_{t_2} := BT_{t_1} \vee ST_{t_1} \vee W_{t_1}$$

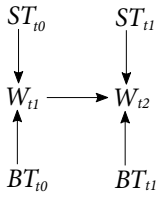
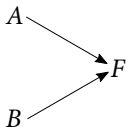



**Figure 7.8** Symmetric Overdetermination

Two arsonists, $A$ and $B$, both throw matches into a gasoline factory. Either match on its own would have sufficed to set the factory on fire, $F$.

$$F := A \vee B$$



(c) But Hitchcock (2001b), Halpern & Pearl (2005), Halpern & Hitchcock (2010), and Weslake (forthcoming) all take it to be a desiderata of an account of causation that it say that both $A$ and $B$ individually caused the factory fire.

23. The case of symmetric overdetermination has motivated the move from *de facto dependence* accounts to what we can call *counterfactual counterfactual dependence* accounts.

### 7.3 Counterfactual Counterfactual Dependence

24. Suppose that we take the basic idea behind De Facto and, instead of merely allowing ourselves to hold variables fixed at their *actual* value, we additionally allow ourselves to hold variables fixed at *counterfactual* values.[12]

    (a) We then get the following general format for a theory of causation:

        Counterfactual Counterfactual
        $C = c$ caused $E = e$ in causal model $\mathcal{M}$ iff

        $$\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models C \neq c \; \square\!\!\rightarrow E \neq e$$

        for suitable variables $\mathbf{G}$ and suitable values $\mathbf{g}$.

    (b) Because the accounts say that causation is counterfactual dependence in some counterfactual scenario, I call these accounts 'counterfactual counterfactual' accounts.

25. All of the accounts in this mold (which, by the way, includes *every* structural theory of causation developed after Hitchcock (2001b)) will agree with Counterfactual Counterfactual. Their differences emerge in the kinds of variables and the kinds of values they treat as suitable.

    (a) This allows us to provide a very straightforward account of the logical relations between these various accounts. If every variable and value that one theory, $T_1$, counts as suitable are counted as suitable by another theory, $T_2$, and $T_2$ counts more variables and values as suitable besides, then we can say that $T_1$ is *strictly weaker* than $T_2$, and we can write '$T_1 \prec T_2$'. (What we mean by '$T_1 \prec T_2$' is just that whatever variable values $T_1$ says are causally related are also said to be causally related by $T_2$, and $T_2$ says some additional variable values are also causally related.)

26. Hitchcock (2001b)'s second theory places the following condition on the counterfactual scenario $\mathbf{G} = \mathbf{g}$: these variable settings should be such that there is some path between $C$ and $E$ such that each variable along this path (besides $C$) takes on its actual value in the counterfactual scenario $\mathbf{G} = \mathbf{g}$.

[12] The presentation in this section owes much to the wonderful taxonomy provided by Weslake (forthcoming).

HITCHCOCK (V2)
**G** and **g** are suitable iff there is some path between $C$ and $E$, $C \to P^1 \to P^2 \to \cdots \to P^k \to E$, such that

$$\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models P^1 = P^1_@ \wedge P^2 = P^2_@ \wedge \cdots \wedge P^k = P^k_@ \wedge E = e$$

(a) This theory says that both $A = 1$ and $B = 1$ are individually causes of $F = 1$ in the symmetric overdetermination case from figure 7.8.

(b) For $A = 1$, we can consider the path $A \to F$, and let $\mathbf{G} = [B]$. Setting $B$ to 0 does not affect the values of any of the variables on the path $A \to F$. That is,

$$\mathcal{M}_{B=0} \models A = 1 \wedge F = 1$$

And both $A$ and $F$ actually take on the value 1. So $B = 0$ is suitable, according to HITCHCOCK (V2).

(c) And, in the counterfactual setting determined by $B = 0$, $F = 1$ counterfactually depends upon $A = 1$.

$$\mathcal{M}_{B=0} \models A = 0 \boxdot\!\!\to F = 0$$

27. Note that the actual setting $\mathbf{G} = \mathbf{G}_@$ of any off-path variables $\mathbf{G}$ will always be suitable according to HITCHCOCK (V2). (If the variables $\mathbf{G} = \mathbf{G}_@$ are not off any path, then they are on every path between $C$ and $E$, in which case $E$ could not counterfactually depend upon $C$.) So, anything that gets ruled as a cause by HITCHCOCK (V1) will get ruled as a cause by HITCHCOCK (V2) as well. (The case of symmetric overdetermination shows us that the converse is false.) So:

$$\text{HITCHCOCK (V1)} \prec \text{HITCHCOCK (V2)}$$

28. HALPERN & PEARL (2005) say that a counterfactual scenario $\mathbf{G} = \mathbf{g}$ is suitable iff, in that counterfactual scenario, $E$ still takes on the value $e$ when we hold fixed $C$ at its actual value $c$—and, moreover, $E$ still takes on the value $e$ when we hold fixed $C$ and *any other* variables which are not in $\mathbf{G}$ at their actual values.

(a) Though they don't build it into their account that there will be some path or paths from $C$ to $E$ such that all the variables in $\mathbf{G}$ are off of those paths, it turns out that such an account would be equivalent to the one they did propose. So we can think of the variables in $\mathbf{G}$ as being the *off-path* variables. And the proposal says that a counterfactual setting of the off-path variables can change the values of variables lying along the path(s), but it can't change the value of $E$, even when we hold fixed $C = c$ and any of the on-path variables at their actual values.

HALPERN AND PEARL
**G** and **g** are suitable iff, for any subvector $\mathbf{G}'$ of $\mathbf{G}$ and any vector of variables $\mathbf{P}$ which are not in $\mathbf{G}$,[13]

$$\mathcal{M}_{\mathbf{G}'=\mathbf{g}} \models (C = c \wedge \mathbf{P} = \mathbf{P}_@) \boxdot\!\!\to E = e$$

[13] I'm writing '$\mathbf{G}' = g$' is mean that the variables in the subvector $\mathbf{G}'$ take on the same values they are assigned by $\mathbf{g}$.

29. If a counterfactual setting $\mathbf{G} = \mathbf{g}$ satisfies the condition set by Hitchcock(v2), then it will definitely satisfy the condition set by Halpern and Pearl (why?). So we can order the various theories we've discussed as follows:

$$\text{Hitchcock (v1)} \prec \text{Hitchcock (v2)} \prec \text{Halpern and Pearl}$$

(a) That is: in any causal model $\mathcal{M}$, if $C = c$ caused $E = e$ according to Hitchcock (v1), then $C = c$ caused $E = e$ according to Hitchcock (v2); and, if $C = c$ caused $E = e$ according to Hitchcock (v2), then $C = c$ caused $E = e$ according to Halpern and Pearl; though none of the converses are true.

## 7.A Exercises

### 7.A.1 Prelude

For easy reference, here are the three structural theories of causation we've discussed:[14]

**Hitchcock (v1)**

According to the causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some vector of variables $\mathbf{G}$ such that

$$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$$

**Hitchcock (v2)**

According to the causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some path between $C$ and $E$, $C \rightarrow P^1 \rightarrow P^2 \rightarrow \cdots \rightarrow P^k \rightarrow E$, some vector of off-path variables $\mathbf{G}$, and an assignment of values $\mathbf{g}$ to those off-path variables such that

1. $\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models P^1 = P^1_@ \wedge P^2 = P^2_@ \wedge \cdots \wedge P^k = P^k_@ \wedge E = e$
2. $\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$

**Halpern and Pearl**

According to the causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some vector of off-path variables $\mathbf{G}$, and an assignment of values $\mathbf{g}$ to those off-path variables such that, for any vector of variables $\mathbf{P}$ not in $\mathbf{G}$ (including the *empty vector*), and any subvector $\mathbf{G}'$ of $\mathbf{G}$,

1. $\mathcal{M}_{\mathbf{G}'=\mathbf{g}} \models (C = c \wedge \mathbf{P} = \mathbf{P}_@) \boxdot\!\!\rightarrow E = e$
2. $\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$

To make it easier to refer to these three theories, let's call them 'H1', 'H2', and 'HP', respectivelly.

Recall that each of these accounts is more liberal than the last—in the sense that each lets in strictly more causes than the one preceding it. That is:

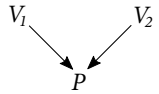
$$\text{H1} \prec \text{H2} \prec \text{HP}$$

Thus, if you can show that $C = c$ caused $E = e$ in $\mathcal{M}$ according to H1, then you've shown that $C = c$ caused $E = e$ in $\mathcal{M}$ according to all three. And, if you can show that $C = c$ *didn't* cause $E = e$ in $\mathcal{M}$ according to HP, then you've shown that $C = c$ *didn't* cause $E = e$ in $\mathcal{M}$ according to all three.

[14] Notice that I've made the choice to interpret '$C \neq c \boxdot\!\!\rightarrow E \neq e$' as meaning that '$(\exists c' \neq c)(C = c' \boxdot\!\!\rightarrow E \neq e)$. That's because some of the exercises will ask you to think about models in which variables can take on more than two values.

### 7.A.2 Exercises

1. Consider the following example.

> **Voting**
>
> A two person committee votes on a proposal. The rules state that, if 50% or more of the committee votes in favor of a proposal, then it passes. So, if the proposal receives at least one vote, it passes. If it receives no votes, it does not pass. Both committee members vote in favor of the proposal, and it passes. Suppose we model this situation with the following variables
>
> $$V_1 = \begin{cases} 0 & \text{if the first member votes against} \\ 1 & \text{if the first member votes for} \end{cases}$$
> $$V_2 = \begin{cases} 0 & \text{if the second member votes against} \\ 1 & \text{if the second member votes for} \end{cases}$$
> $$P = \begin{cases} 0 & \text{if the proposal doesn't pass} \\ 1 & \text{if the proposal passes} \end{cases}$$
>
> (the first two variables are exogenous, and the last endogenous) and the following structural equation:
>
> $$\left[\ P := V_1 \vee V_2\ \right]$$
>
> 
> 
>
> (where $X \vee Y = \max\{X, Y\}$). In the actual context, $V_1 = V_2 = 1$.

   (a) Before answering the questions below: in your opinion, did the first member's vote cause the proposal to pass? (For all questions where I explicitly ask for your opinion, there is no uniquely correct answer; only honest and dishonest answers.)

   (b) Does H1 say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

   (c) Does H2 say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

   (d) Does HP say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

2. Consider the following example.

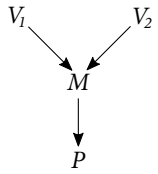
VOTING MACHINE (HALPERN & PEARL, 2005, p. 881)

A two person committee votes on a proposal. The rules state that, if 50% or more of the committee votes in favor of a proposal, then it passes. So, if the proposal receives at least one vote, it passes. If it receives no votes, it does not pass. The voting works like this: each member feeds their votes into a machine which calculates the total number of votes in favor. If the machine reads '1' or above after voting, then the proposal passes. Both committee members vote in favor of the proposal, and it passes.
Suppose we model this situation with the following variables

$$V_1 = \begin{cases} 0 & \text{if the first member votes against} \\ 1 & \text{if the first member votes for} \end{cases}$$
$$V_2 = \begin{cases} 0 & \text{if the second member votes against} \\ 1 & \text{if the second member votes for} \end{cases}$$
$$M = \begin{cases} 0 & \text{if the machine reads '0'} \\ 1 & \text{if the machine reads '1'} \\ 2 & \text{if the machine reads '2'} \end{cases}$$
$$P = \begin{cases} 0 & \text{if the proposal doesn't pass} \\ 1 & \text{if the proposal passes} \end{cases}$$

(the first two variables are exogenous, and the last endogenous) and the following structural equations:

$$\left[ \begin{array}{l} M := V_1 + V_2 \\ P := \begin{cases} 1, \text{ if } M \geq 1 \\ 0, \text{ if } M = 0 \end{cases} \end{array} \right]$$




In the actual context, $V_1 = V_2 = 1$.

(a) Before answering the questions below: in your opinion, did the first member's vote cause the proposal to pass?

(b) Does H1 say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

(c) Does H2 say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

(d) Does HP say that, in this model, $V_1 = 1$ is a cause of $P = 1$?

(e) Did any of the theories give different answers in VOTING MACHINE and VOTING?

3. Consider the following example.

LIGHT SWITCH (PEARL, 2000, p. 324)

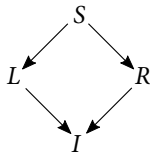
There is a switch which can be flipped to either the left or the right. Iff the switch is flipped to the left, then a lamp on the left will be turned on. Iff the switch is flipped to the right, then a lamp on the right will be turned on. Iff either lamp is on, then the room will be illuminated. The switch is flipped to the left, and the room is illuminated.

Suppose we model this situation with the following variables:

$$S = \begin{cases} 0 & \text{if the switch is to the left} \\ 1 & \text{if the switch is to the right} \end{cases}$$
$$L = \begin{cases} 0 & \text{if the left lamp is off} \\ 1 & \text{if the left lamp is on} \end{cases}$$
$$R = \begin{cases} 0 & \text{if the right lamp is off} \\ 1 & \text{if the right lamp is on} \end{cases}$$
$$I = \begin{cases} 0 & \text{if the room is not illuminated} \\ 1 & \text{if the room is illuminated} \end{cases}$$

(the first variable is exogenous, and the others endogenous) and the following structural equations:

$$\begin{bmatrix} L := \neg S \\ R := S \\ I := L \vee R \end{bmatrix}$$




(where $X \vee Y = \max\{X, Y\}$ and $\neg X = 1 - X$). In the actual context, $S = 0$.

(a) Before answering the questions below: in your opinion, did the switch's being set to the left cause the room to be illuminated?

(b) Does H1 say that, in this model, $S = 0$ is a cause of $I = 1$?

(c) Does H2 say that, in this model, $S = 0$ is a cause of $I = 1$?

(d) Does HP say that, in this model, $S = 0$ is a cause of $I = 1$?

4. Consider the following example.

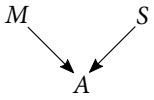
MAJOR/SERGEANT

The Major outranks the Sergeant, and the soldiers always follow the order of the highest-ranking officer. Both the Major and the Sergeant order the soldiers to advance. They advance.
Suppose we model this situation with the following variables

$$M = \begin{cases} 0 & \text{if the major gives no order} \\ 1 & \text{if the major orders to advance} \end{cases}$$
$$S = \begin{cases} 0 & \text{if the sergeant gives no order} \\ 1 & \text{if the sergeant order to advance} \end{cases}$$
$$A = \begin{cases} 0 & \text{if the soldiers don't advance} \\ 1 & \text{if the soldiers advance} \end{cases}$$

(the first two variables are exogenous, and the last endogenous) and the following structural equations:

$$\left[\ A := M \vee S\ \right]$$




In the actual context, $M = S = 1$.

(a) i. Before answering the questions below: in your opinion, did the Major's order cause the soldiers to advance?
   ii. Before answering the questions below: in your opinion, did the Sergeant's order cause the soldiers to advance?

(b) i. Does H1 say that, in this model, $M = 1$ is a cause of $A = 1$?
   ii. Does H1 say that, in this model, $S = 1$ is a cause of $A = 1$?

(c) i. Does H2 say that, in this model, $M = 1$ is a cause of $A = 1$?
   ii. Does H2 say that, in this model, $S = 1$ is a cause of $A = 1$?

(d) i. Does HP say that, in this model, $M = 1$ is a cause of $A = 1$?
   ii. Does HP say that, in this model, $S = 1$ is a cause of $A = 1$?

5. Consider the following example.

MAJOR/SERGEANT (V2)

The Major outranks the Sergeant, and the soldiers always follow the order of the highest-ranking officer. Both the Major and the Sergeant order the soldiers to advance. They advance.

Suppose we model this situation with the following variables

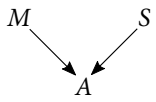
$$M = \begin{cases} -1 & \text{if the Major orders to retreat} \\ 0 & \text{if the Major gives no order} \\ 1 & \text{if the Major orders to advance} \end{cases}$$

$$S = \begin{cases} -1 & \text{if the Sergeant orders to retreat} \\ 0 & \text{if the Sergeant gives no order} \\ 1 & \text{if the Sergeant orders to advance} \end{cases}$$

$$A = \begin{cases} -1 & \text{if the soldiers retreat} \\ 0 & \text{if the soldiers remain in place} \\ 1 & \text{if the soldiers advance} \end{cases}$$

(the first two variables are exogenous, and the last endogenous) and the following structural equations:

$$\left[ A := \begin{cases} S & \text{if } M = 0 \\ M & \text{if } M \neq 0 \end{cases} \right]$$



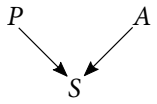

In the actual context, $M = S = 1$.

(a) i. Does H1 say that, in this model, $M = 1$ is a cause of $A = 1$?
ii. Does H1 say that, in this model, $S = 1$ is a cause of $A = 1$?

(b) i. Does H2 say that, in this model, $M = 1$ is a cause of $A = 1$?
ii. Does H2 say that, in this model, $S = 1$ is a cause of $A = 1$?

(c) i. Does HP say that, in this model, $M = 1$ is a cause of $A = 1$?
ii. Does HP say that, in this model, $S = 1$ is a cause of $A = 1$?

6. Consider the following example.

> Bogus Prevention (Hiddleston, 2005, p. 32)
> Margaery has orders from Olenna to place the poison in Joffrey's cup. However, at the last minute, she has a change of heart and does not pour the poison in the cup. As it turns out, Jaime was aware that Joffrey's cup might be poisoned, so he had placed a harmless antidote in the cup. had Margaery poured the poison, the antidote would have neutralized it, and Joffrey would have survived.
> Suppose we model this situation with the following variables (***pay special attention to the definition of*** $P$)
>
> $$P = \begin{cases} 0 & \text{if Margaery pours the poison into Joffrey's cup} \\ 1 & \text{if Margaery doesn't pour the poison into Joffrey's cup} \end{cases}$$
> $$A = \begin{cases} 0 & \text{if Jaime doesn't pour the antidote intp Joffrey's cup} \\ 1 & \text{if Jaime pours the antidote into Joffrey's cup} \end{cases}$$
> $$S = \begin{cases} 0 & \text{if Joffrey doesn't survive (i.e., if he is dead at the end of the day)} \\ 1 & \text{if Joffrey survives (i.e., if he is alive at the end of the day)} \end{cases}$$
>
> (the first two variables are exogenous, and the last endogenous) and the following structural equations:
>
> $$\left[\ S := P \vee A\ \right]$$
>
> 
> 
>
> In the actual context, $P = A = 1$.

   (a) Before answering the questions below: in your opinion, did Jaime's pouring the antidote into Joffrey's cup cause him to be alive at the end of the day?

   (b) Does H1 say that, in this model, $A = 1$ is a cause of $S = 1$?

   (c) Does H2 say that, in this model, $A = 1$ is a cause of $S = 1$?

   (d) Does HP say that, in this model, $A = 1$ is a cause of $S = 1$?

7. What's interesting about the structural equations models from Questions 1 and 6?

# 8 Normality

## 8.1 Troubles with Underdetermination

1. Where we're at in the course:

   (a) Counterfactual theories of causation ran into troubles with cases of preemption and trumping.

   (b) Lewis (2000)'s 'influence' account was an attempt to deal with these troubles, but we saw reasons to believe that it ultimately lets in far too many causes.

   (c) We considered Yablo (2002, 2004)'s proposal that causation is (or is in some way related to) 'de facto dependence'—that is, counterfactual dependence, holding fixed some true fact about the world $G$ (the *grounds* of the causal relation).

   (d) We saw that this approach could be implemented within the framework of *structural equations models* (or, as they are often called 'causal models').

   (e) In that framework, the account says:

   > De Facto
   > In causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some suitable vector of variables $\mathbf{G}$, with actual values $\mathbf{G}_@$, such that[1]
   >
   > $$\mathcal{M}_{\mathbf{G}=\mathbf{G}_@} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$$

   (f) We also saw that several authors take cases of symmetric overdetermination (e.g., the neuron diagram in figure 8.1(a)) as reasons to weaken De Facto so that it lets in more causes. We called these accounts 'counterfactual counterfactual dependence' accounts, since they say that causation is not *actual*, or even *de facto* counterfactual dependence, but rather counterfactual dependence *in some counterfactual scenario*.

   > Counterfactual Counterfactual
   > In causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some suitable vector of variables $\mathbf{G}$, with suitable values $\mathbf{g}$, such that
   >
   > $$\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$$

[1] Here and below, I've made the choice to interpret '$C \neq c \boxdot\!\!\rightarrow E \neq e$' as being true when and only when there is some value of $C$, $c'$, such that had $C$ taken on the value $c'$, $E$ would not have taken on the value $e$. Remember that there are alternative choices available.

2. I think it's appropriate to view both De Facto and Counterfactual Counterfactual as descendants of Lewis's counterfactual theory of causation.

   (a) Both these accounts attempt to understand causation using *only* counterfactual information. The defect they see in earlier counterfactual theories is just that they did not include *enough* counterfactual information.

   (b) The hope of De Facto and Counterfactual Counterfactual is that, by paying attention to the richer counterfactual structure encoded in a correct structural equations model, we can give a more satisfying treatment of causation.

3. What we saw in the exercises from last time, however, is an argument that even this richer counterfactual structure is insufficient to properly characterize singular causal relations.

   (a) To illustrate, let's consider the two neuron diagrams from figure 8.1.

   (b) The case of symmetric overdetermination (figure 8.1(a)) is controversial. Some (*e.g.*, Lewis and Mackie) think that intuitions are murky, and it's no great cost to say that $A$'s firing didn't cause $C$'s firing. Others, like Hitchcock (2001b) and Halpern & Pearl (2005), think that a theory of causation ought to say that $A$'s firing caused $C$'s firing.

   (c) The case of Bogus Prevention (figure 8.1(b)), however, is uncontroversial. $a$'s firing did not prevent $c$ from firing (that is to say, $a$'s firing did not cause $c$ to not fire).

   (d) Note that we can model the case of symmetric overdetermination from figure 8.1(a) with the following variables:

$$A = \begin{cases} \mathbf{1} & \text{if neuron } A \text{ fires at } t_1 \\ 0 & \text{if neuron } A \text{ doesn't fire at } t_1 \end{cases}$$
$$B = \begin{cases} \mathbf{1} & \text{if neuron } B \text{ fires at } t_1 \\ 0 & \text{if neuron } B \text{ doesn't fire at } t_1 \end{cases}$$
$$C = \begin{cases} \mathbf{1} & \text{if neuron } C \text{ fires at } t_2 \\ 0 & \text{if neuron } C \text{ doesn't fire at } t_2 \end{cases}$$

   (the actual values of the variables are bolded) and the following structural equation:

$$C := A \vee B$$

   Call this model '$\mathcal{M}_{so}$'.

   (e) And note that we can model the case of Bogus Prevention from figure 8.1(b) with the following variables:

**Figure 8.1** A pair of counterfactually isomorphic neuron diagrams


(a) Symmetric Overdetermination (b) Bogus Prevention

$$a = \begin{cases} \mathbf{1} & \text{if neuron } a \text{ fires at } t_1 \\ 0 & \text{if neuron } a \text{ doesn't fire at } t_1 \end{cases}$$
$$b^* = \begin{cases} 0 & \text{if neuron } b \text{ fires at } t_1 \\ \mathbf{1} & \text{if neuron } b \text{ doesn't fire at } t_1 \end{cases}$$
$$c^* = \begin{cases} 0 & \text{if neuron } c \text{ fires at } t_2 \\ \mathbf{1} & \text{if neuron } c \text{ doesn't fire at } t_2 \end{cases}$$

(actual variable values are bolded) and the following structural equation:

$$c^* := a \vee b^*$$

Call this model '$\mathcal{M}_{bp}$'.

(f) These two structural equations models are isomorphic to each other. If we take the first model and just swap out '$A$' with '$a$', '$B$' with '$b^*$', and '$C$' with '$c^*$', we get back the second model. And, *moreover*, the actual value of $A$ matches the actual value of $a$; the actual value of $B$ matches the actual value of $b^*$, and the actual value of $C$ matches the actual value of $c^*$:

$$A = a = 1$$
$$B = b^* = 1$$
$$C = c^* = 1$$

(g) We could then make the following argument:

P1. $A$'s firing caused $C$ to fire.

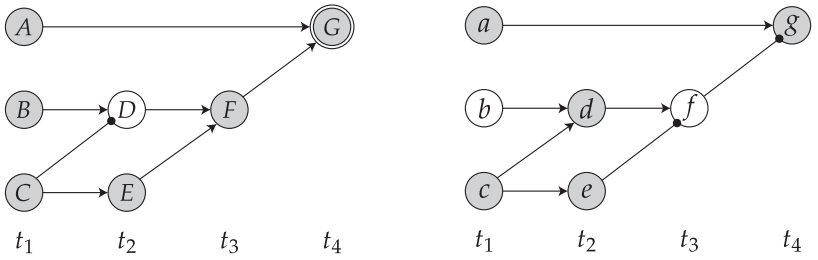
P2. $a$'s firing did not cause $c$ to not fire.

P3. Both of $\mathcal{M}_{so}$ and $\mathcal{M}_{bp}$ are correct.

---

C1. A correct structural equations model underdetermines singular causal relations.

**Figure 8.2** A second pair of counterfactually isomorphic neuron diagrams. (In figure 8.2(a), the neuron $G$ is a 'dull' neuron. It requires two stimulations to fire. Thus, it will only fire if both $A$ and $F$ fire.)




(a) Early Preemption (b) A 'Short Circuit'

4. This argument will be convincing to those—like Hitchcock and Halpern & Pearl—who are convinced of P1. However, those—like Mackie and Lewis—who were already on the fence about P1, and were happy to accept theories of causation which rejected it, may be more inclined to respond by saying that cases like *Bogus Prevention* have simply afforded us an argument that, in fact, $A$'s firing *didn't* cause $C$'s firing.

5. I don't think, however, that matters are so easily settled. For there are other cases of isomorphic structural equations models where the causal judgments are less murky.

   (a) Consider the pair of neuron diagrams shown in figure 8.2. (This is the original example of counterfactually isomorphic neuron diagrams provided in Hall (2007); though, ever since, the literature has focused on examples like figure 8.1.)

   (b) It is uncontroversial that, in figure 8.2(a), $C$'s firing caused $G$'s firing.

   (c) We first encountered the structure in figure 8.2(b) in Lewis (2004)'s discussion of transitivity (figure 3.4, p. 98).

      i. There, the neurons $c$, $e$, and $f$ provide an apparent counterexample to transitivity: for $c$'s firing caused $e$ to fire; and $e$'s firing prevented $f$ from firing; but it does not appear that $c$'s firing prevented $f$ from firing.

      ii. Figure 8.2(b) hooks this apparent violation of transitivity up with a case of double prevention. By preventing $f$ from preventing $g$ from firing, $e$ caused $g$ to fire. However, again, it does not appear that $c$'s firing caused $g$ to fire.

   (d) However, just as with the pair of neuron diagrams in figure 8.1, the neuron diagrams in figures 8.2(a) and 8.2(b) may be modeled with *isomorphic* structural equations models (that is to say: their counterfactual structure is isomorphic).

For the neuron diagram in figure 8.2(a), we have the equations

$$\begin{bmatrix} G := A \wedge F \\ F := D \vee E \\ D := B \wedge \neg C \\ E := C \end{bmatrix}$$

(where the variables have the natural interpretation), and we have the actual values

$$A = B = C = E = F = G = 1, \qquad D = 0$$

Call this model '$\mathcal{M}_{ep}$'.

(e) For the neuron diagram in figure 8.2(b), we have the isomorphic equations

$$\begin{bmatrix} g := a \wedge f^* \\ f^* := d^* \vee e \\ d^* := b^* \wedge \neg c \\ e := c \end{bmatrix}$$

(where the variables have the natural interpretation, except for the $*$-ed variables, which have the value 0 if their associated neurons fire at their designated time and have the value 1 if their associated neurons *don't* fire at their designated time), and we have the actual values

$$a = b^* = c = e = f^* = g = 1, \qquad d^* = 0$$

Call this model '$\mathcal{M}_{sc}$'.

(f) We can then make the following argument:

P1. $C$'s firing caused $G$ to fire.

P2. $c$'s firing did not cause $g$ to not fire.

P3. Both of $\mathcal{M}_{ep}$ and $\mathcal{M}_{sc}$ are correct.

---

C1. A correct structural equations model underdetermines singular causal relations.

6. Lewis (2000) rejects P2 (though he confesses to feeling "some ambivalence"). For those of us who accept it, however, we are forced to abandon the search for a theory of causation formulated in terms of structural equations models.

7. In fact, something more radical follows. For the neuron diagrams in figures 8.1 and 8.2, the structural equations models provide all the counterfactual information there is to have. If this information isn't enough to tell us which variable values are causally related, then *no amount* of counterfactual information is enough to tell us which variable values are causally related.

P1. $C$'s firing caused $G$ to fire.

P2. $c$'s firing did not cause $g$ to not fire.

P3. $\mathcal{M}_{ep}$ and $\mathcal{M}_{sc}$ entail all and only the true counterfactuals about the neuron diagrams from figures 8.2(a) and 8.2(b).

---

C1. Singular causal relations do not supervene upon counterfactual facts alone.

## 8.2 Omissions, Normality, and Normativity

8. Both of the examples we've seen so far involve omissions, or negative facts/events.

   (a) In figure 8.1(b), both $b$ and $c$ *failed to fire*; whereas, in figure 8.1(a), $B$ and $C$ fired.

   (b) In figure 8.2(b), $f$ *failed to fire*; whereas, in figure 8.2(a), $F$ fired.

9. McGrath (2005) gives further evidence that, at least when we're considering cases in which an omission occurs in cause-position, counterfactual facts alone do not determine singular causal facts.

10. She asks us to consider the following case:

    **Plant** Barry promises his neighbor Alice that he will water her plant while she is away on vacation. However, Barry does not water the plant and the plant becomes dehydrated and dies. Alice has another neighbor, Carlos, who also does not water the plant.

    (a) In Plant, it appears that

    Barry Barry's failure to water the plant caused it to die.

    is true. However, it appears that

    Carlos # Carlos's failure to water the plant caused it to die.

    is false.

11. McGrath's view is that, in Plant, appearances are not misleading. It is literally true that Barry's failure to water the plant caused it to die, and it is literally false that Carlos's failure to water the plant caused it to die.

    (a) If we agree, then we have yet another argument that counterfactual facts underdetermine causal facts; for all the relevant counterfactuals which are true of Barry are true of Carlos as well.

    (b) Moreover, all the counterfactual accounts of causation we've considered accept that counterfactual dependence is sufficient for causation. But the plant's death counterfactually depends upon Carlos's failure to water it.

12. Three (non-exhaustive) positions:

    (a) The *semantic* position: Barry is true and Carlos is false.

    (b) The *pragmatic* position: while both Barry and Carlos are true, it is pragmatically inappropriate to utter Carlos.

    (c) The *anti-omissions* position: omissions cannot be causes or effects, so neither Barry nor Carlos is true. Nevertheless, it is pragmatically appropriate to assert Barry, and pragmatically inappropriate to assert Carlos.

13. (a) If we adopt (12a), then we need to say something about what the difference is between Barry and Carlos—why does Barry's omission cause the plant to die, but not Carlos's?

        i. McGrath (2005) opts for option (12a), and argues that our accounts of causation must be sensitive, at least in some cases, to *normative* facts.

    (b) If we adopt (12b) or (12c), then we need to give a general pragmatic account which explains why Barry is assertible and Carlos is not.

14. The anti-omissions position may be correct, but it is not sufficient to handle all cases in which normativity seems to infect our causal judgments. For consider a case adapted from Hitchcock & Knobe (2009):

    **Pens** Due to a recent pen shortage, administrators are allowed to take pens from the pen jar, but professors are not. Professor Procrastinate and Administrator Angelina run into each other in the office. Procrastinate is scrambling to get his referee report done, and Angelina is filling out reimbursement forms. Both take pens from the pen jar. Later, Administrator Albert needs a pen to take down an important message from the dean, but there are no pens left. Albert is fired for failing to take down the message from the dean.

    (a) It seems appropriate to say

        Procrastinate Procrastinate's taking a pen caused Albert to get fired.

        but inappropriate to say:

        Angelina Angelina's taking a pen caused Albert to get fired.

15. McGrath (2005) additionally argues against the pragmatic account. She offers two challenges such an account would have to surmount.

    (a) Firstly, Gricean mechanisms in general explain why uttering $\ulcorner p \urcorner$ would be infelicitous by pointing out that $\ulcorner p \urcorner$ implicates that $q$, and $q$ is false.

        i. Thus, for instance, saying "Some of the boys went to the lake" implicates that not all of the boys went. So, if all of the boys went, then it is infelicitous to utter "Some of the boys went to the lake".

    (b) However, just because $\ulcorner p \urcorner$ implicates $q$, this does not mean that $\ulcorner \neg p \urcorner$ implicates $\neg q$.

i. For instance, "None of the boys went to the lake" doesn't implicate that all of the boys went.

but the data isn't just that we hesitate to assert Carlos. We are willing to assert its negation, as well.

(c) Secondly, the data to be explained isn't just data about our *assertions*. We additionally *believe* that Carlos is false. Pragmatics doesn't provide a persuasive error theory of this *belief*.

16. McGrath reaches the conclusion, ultimately, that standards of *normality* have a role to play in determining which events are causally related.

(a) *Normality* is a broad notion which can include moral normativity, statistical regularity, and proper functioning.

(b) McGrath gives a non-reductive account of the relationship between causation and normality in the special case involving omissions in cause position. That proposal is:

> McGrath (2005)
> An omission $o$ caused a positive event $e$ iff both $o$ and $e$ occur, and either:
>
> i. the kind of event of which $o$ is the omission is a *normal would-be preventer* of $e$ (if $e$ were to be prevented, it would have been normal for an event of this type to prevent it); or
>
> ii. the kind of event of which $o$ is the omission is a *normal would-be preventer* of $e^*$, and $e^*$ caused $e$.

17. Similar conclusions are reached, on independent grounds, by Maudlin (2004). Maudlin argues that, when we make causal claims, we conceive of a situation as (what he calls) a 'quasi-Newtonian system'. A quasi-Newtonian system is one in which things have certain *default, inertial* states that they will remain in unless acted upon, in which case they will enter into *deviant, non-inertial* states.

(a) Wedding the accounts, we could think that *normal* behavior is default, or inertial, while *abnormal* behavior is deviant, or non-inertial.

18. It would be nice to incorporate these notions into a full account of causation.

## 8.3 Incorporating Normality

19. Notice that we could appeal to Mackie's notion of a *causal field* to try to explain why Barry and Procrastinate are (in context) true, yet Carlos and Angelina are (in context) false.

(a) Recall, Mackie thought that causal claims were made and evaluated relative to a *causal field*. In the case of singular causation, the causal field consists of all situations similar to the one we are considering.

(b) The causal field will generate a list of all features of the situation in question which we are taking for granted. Thus, e.g., if no situation in the causal field is one without oxygen present, then we are taking for granted the presence of oxygen.

(c) On Mackie's theory, features taken for granted are not felicitously cited as causes (this is due to the constraint that a cause be a part of a *minimally* sufficient condition for the effect). So, if our causal field is one which takes for granted the presence of oxygen, then

> The presence of oxygen caused the match to light.

will be false, relative to that causal field.

(d) Suppose that, in general, when evaluating singular causal or explanatory claims, we take for granted those features of the case under consideration which are *normal* (in McGrath's sense). This would mean that normal features of the case are not appropriately cited as causes.

(e) Then, in Plant, if we suppose that it is *normal* that Carlos didn't water the plant, Mackie's account would tell us that "Carlos's failure to water the plant caused it to die" is, in ordinary contexts, false.

(f) And, in Pens, if we suppose that it is *normal* for Angelina to take a pen, Mackie's account would tell us that "Angelina's taking a pen caused Albert to get fired" is, in ordinary contexts, false.

20. Several authors attempt to incorporate normality into counterfactual counterfactual accounts.[2] The general strategy is to say that the only counterfactual settings $\mathbf{G} = \mathbf{g}$ which are appropriate for 'unmasking' the causal relationship between $C = c$ and $E = e$ are those in which the variable values in $\mathcal{M}_{\mathbf{G}=\mathbf{g}}$ are *at least as normal* as those in $\mathcal{M}$.

21. For instance, here is the account from Halpern (2008) (which adds an additional normality requirement to the definition of Halpern & Pearl (2005)):

> Halpern (2008)
> In causal model $\mathcal{M}$, $C = c$ caused $E = e$ iff there is some vector of off-path variables $\mathbf{G}$ and some vector of values $\mathbf{g}$ such that, for any vector of variables $\mathbf{P}$ not in $\mathbf{G}$ (including the *empty vector*), and any subvector $\mathbf{G}'$ of $\mathbf{G}$ (including the *empty vector*):
> (a) $\mathcal{M}_{\mathbf{G}=\mathbf{g}} \models (\exists c')(C = c' \boxdot\!\!\rightarrow E \neq e)$
> (b) $\mathcal{M}_{\mathbf{G}'=\mathbf{g}} \models (C = c \wedge \mathbf{P} = \mathbf{P}_{@}) \boxdot\!\!\rightarrow E = e$
> (c) The variable values taken on in the counterfactual model $\mathcal{M}_{\mathbf{G}=\mathbf{g}}$ are *at least as normal as* the variables values taken on in the actual model $\mathcal{M}$.

[2] Besides Halpern (2008), see Hitchcock (2007) and Hall (2007).

## 8.A Exercises

Making the assumption that it is more normal for neurons to not fire than for them to fire (and that, therefore $\mathcal{M}_{\mathbf{G}=\mathbf{g}}$ is at least as normal as $\mathcal{M}$ iff every neuron which fires in $\mathcal{M}_{\mathbf{G}=\mathbf{g}}$ also fires in $\mathcal{M}$), how does Halpern (2008) answer the following questions:

1. In figure 8.1(a), does $A$'s firing cause $C$'s firing (according to the causal model $\mathcal{M}_{so}$)?
2. In figure 8.1(b), does $a$'s firing cause $c$'s failure to fire (according to the causal model $\mathcal{M}_{bp}$)?
3. In figure 8.2(a), does $C$'s firing cause $G$'s firing (according to the causal model $\mathcal{M}_{ep}$)?
4. In figure 8.2(b), does $c$'s firing cause $g$'s firing (according to the causal model $\mathcal{M}_{sc}$)?

# 9 Interventionist Theories

## 9.1 Causation and Agency

1. The *Agency* theory of causation—or, as it is alternatively called, the *Manipulationist*, or *Interventionist* theory of causation—contends that our experience as agents causally impacting the world around us forms an integral part of our concept of causation; and, furthermore, that causation can be analyzed in terms of agency, manipulation, or intervention.

   (a) A bare-bones, preliminary version of the account: $c$ caused $e$ iff bringing about $c$ would be an effective means for an agent to bring about $e$.

2. Menzies & Price (1993) put meat on the bones of this account by specifying what it is that makes $c$ an "effective means" for bringing about $e$.

   (a) Their account begins with the idea of an *agent probability* for $e$, given that you have brought about $c$ as the result of a free action—$\Pr_c(e)$. We are told that these are the probabilities which ought to enter into our theory of rational action.

   (b) The theories of rational action Menzies & Price are thinking of follow the same basic outline:

      i. Firstly, any well-formulated decision problem will include a specification of the various possible *actions* available to you, $a_1, a_2, \ldots, a_N$,[1] the various possible *states of the world* which might obtain, $s_1, s_2, \ldots, s_M$ (where these states of the world form a partition),[2] and a *value function* $V$, defined over conjunctions of acts and states.

      ii. The theories state that an action $a_i$ is rational iff it maximizes the function $EV$ (defined over acts), where

      $$EV(a_i) = \sum_j \Pr_{a_i}(s_j) \cdot V(a_i \wedge s_j)$$

      iii. For so-called 'evidential decision theorists', $\Pr_a(s)$ is just the conditional probability $\Pr(s \mid a)$.

[1] The $a_i$'s are, strictly speaking, not the actions themselves, but rather the propositions that you perform those various actions.

[2] Again, these are the propositions that those states of the world obtain.

iv. For so-called 'causal decision theorists' (or, at least, for Gibbard & Harper (1981)'s versions of that theory), $\Pr_a(s)$ is the probability of the counterfactual 'were you do to $a$, $s$ would result', $\Pr(a \mathrel{\square\!\!\rightarrow} s)$.[3]

(c) Given this basic theory of rational action, Menzies & Price say that we should think of $c$ as an effective means for bringing about $e$ iff: were you in a position to either bring about $c$ or $\neg c$ as the result of a free action, then, were you to value $e$ and $e$ alone, it would be rational for you to bring about $c$.

i. That means that $c$ is an effective means for bringing about $e$ iff:[4]

$$\Pr_c(e) \cdot V(e) + \Pr_c(\neg e) \cdot V(\neg e) > \Pr_{\neg c}(e) \cdot V(e) + \Pr_{\neg c}(\neg e) \cdot V(\neg e)$$

ii. We can show that, if we suppose that $V(e) > V(\neg e)$, then the above inequality holds iff $\Pr_c(e) > \Pr_{\neg c}(e)$. For the above may be re-written as:

$$\Pr_c(e)\,[V(e) - V(\neg e)] > \Pr_{\neg c}(e)\,[V(e) - V(\neg e)]$$

And, since $V(e) - V(\neg e)$ is positive, we may divide both sides by it, and we get $\Pr_c(e) > \Pr_{\neg c}(e)$.

3. Thus, Menzies & Price (1993)'s interventionist account of causation says the following:

> Menzies & Price
> The token event $c$ caused the token event $e$ iff:
>
> $$\Pr_c(e) > \Pr_{\neg c}(e)$$

4. Though this looks superficially like the probabilistic theories we studied before, the appeal to agent probabilities allows it to evade many of the problem cases which plagued those accounts.

(a) Recall, one problem case was provided by joint effects of a common cause, as in:

> Barometer
> The atmospheric pressure causes both the barometer reading and the storm. For this reason, the probability that there is a storm, given that the barometer reading is low, $\Pr(s \mid b)$, is higher than the probability that there is a storm, given that the barometer reading is high, $\Pr(s \mid \neg b)$. Nevertheless, the barometer reading does not cause the storm.

i. While *conditionalizing* on the barometer reading does raise the probability of a storm, *intervening* so as to raise the barometer reading does nothing

[3] There are several different versions of causal decision theory; see, for starters, Lewis (1981), Skyrms (1980b), Joyce (1999), and Meek & Glymour (1994).

[4] We have '$V(e)$' and '$V(\neg e)$', rather than '$V(c \wedge e)$' and the like, because we have stipulated that you care about $e$ and $e$ alone, so $V(c \wedge e) = V(\neg c \wedge e) = V(e)$.

to affect the probability of a storm—that is, $\Pr_b(s) = \Pr_{\neg b}(s)$. Imagine that you were able to reach into the barometer and raise the mercury level. Doing so would not affect the probability of a storm at all. So Menzies & Price do not say that the barometer reading caused the storm.

(b) Recall, another problem for those theories was the symmetry of probabilistic dependence. That is, whenever $\Pr(e \mid c) > \Pr(e \mid \neg c)$, $\Pr(c \mid e) > \Pr(c \mid \neg e)$. However, causation is not symmetric. Probabilistic theorists got around this by stipulating the causes must precede their effects.

i. Menzies & Price need not appeal to the direction of time to fix the direction of causation. They can simply note that, while *conditional* probability raising is symmetric, *interventionist* probability raising is not.

ii. Note that the agent probability of a low barometer reading, given that you've raised the atmospheric pressure, is high. That is, if you imagine yourself capable of bringing about high atmospheric pressure through a voluntary act, doing so would raise the probability of a low barometer reading. So $\Pr_a(b) > \Pr_{\neg a}(b)$. However, if you imagine yourself capable of bringing about a low barometer reading through a voluntary act, doing so would not raise the probability of high atmospheric pressure. So $\Pr_b(a) = \Pr_{\neg b}(a)$.

5. It's worth noting that much of the content of this view depends upon the precise way in which the *agent probabilities* $\Pr_c(e)$ are understood.

(a) For instance, if we understand them as Gibbard & Harper understand them, then the account says that $c$ caused $e$ iff $\Pr(c \boxrightarrow e) > \Pr(\neg c \boxrightarrow e)$. In deterministic cases of preemption, this would rule that neither the preempting cause nor the preempted backup are causes of the effect (more on this below).

6. Though Menzies & Price do not suggest this themselves—the suggestion actually comes from Meek & Glymour (1994)—we could alternatively understand agent probabilities within the framework of *causal Bayes nets*.

(a) This framework will allow us to rigorously draw the distinction between *conditioning* and *intervening*.

(b) The framework will also set the stage for Woodward (2003)'s interventionist theory of causation.

(c) It will additionally introduce us to the *Markov condition*, which we will be discussing in further depth later on in the course.

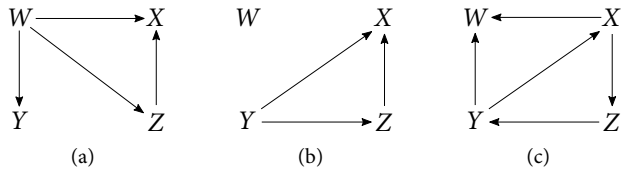
## 9.2 Causal Bayes Nets

7. A *causal Bayes net* $\mathcal{B}$ is a triple $< \mathcal{V}, \mathcal{E}, \Pr >$, where:

(a) $\mathcal{V}$ is a set of variables $\{V_1, V_2, \ldots, V_N\}$;

(b) $\mathcal{E}$ is a set of (acyclic) *directed edges* between the variables in $\mathcal{V}$; and

**Figure 9.1** Possible directed edges, $\mathcal{E}$, for the variable set $\mathcal{V} = \{W, X, Y, Z\}$. While the directed edges in figures 9.1(a) and 9.1(b) are acyclic, those in figure 9.1(c) are not.




(c) Pr is a joint probability distribution Pr defined over the values of the variables in $\mathcal{V}$ with the following property (this property is known as the *Markov condition*: for any variables $X, Y$ such that $Y \notin \mathbf{DE}(X)$,

$$\Pr(X \mid Y, \mathbf{PA}(X)) = \Pr(X \mid \mathbf{PA}(X)) \qquad \text{(MC)}$$

That is: so long as $Y$ is not a causal descendant of $X$, $X$ and $Y$ are probabilistically independent, once we conditionalize on $X$'s causal parents.

i. Directed edges and the Markov condition shouldn't make sense just yet—but we'll explain them below.

8. Variables are to be understood just as they were in our discussion of structural equations models.

(a) Variables are functions from possibilities to numbers on the real line, which are understood to represent some determinable property of the world; the values of that variable are its determinants.

(b) We write '$V_w = v$' to say that the value of the world $w$, under the function $V$, is $v$—or, more intuitively, that at $w$, the determinant of the determinable $V$ is $v$.

(c) Given a variable $V$, one of whose values is $v$, we can form the proposition $V = v$. This is just the proposition that the value of $V$ is $v$. It is the set of possibilities $w$ such that $V_w = v$. $V = v \stackrel{\text{def}}{=} \{w \mid V_w = v\}$. These are the propositions to which Pr will be assigning probabilities.

9. *Directed edges* could be defined as ordered pairs of variables from $\mathcal{V}$, but they are most naturally thought of in terms of *graphs*. For instance, if we have the variable set $\mathcal{V} = \{W, X, Y, Z\}$, three sets of directed edges are shown in figure 9.1.

(a) Not just any set of directed edges are allowed in a causal Bayes net. We explicitly forbid those like the one in figure 9.1(c)—those which contain *loops*, or *cycles*. That is: in a causal Bayes net, we only permit *acyclic* directed edges.

(b) Some definitions:

i. The set of $V$'s *causal parents*, $\mathbf{PA}(V)$, is the set of all those variables $X$ such that there is a directed edge leading from $X$ to $V$, $X \rightarrow V$.

ii. The set of $V$'s *causal descendants*, $\mathbf{DE}(V)$, is the set of all those variables $X$ such that you can reach $X$ from $V$ by following directed edges tail-to-tip—that is, $X \in \mathbf{DE}(V)$ iff there is some sequence of intermediate variables $I_1, \ldots, I_N$ such that $V \to I_1 \to \cdots \to I_N \to X$. (Additionally: we stipulate, somewhat unnaturally, that every variable is one of its own descendants; that is, for all $V \in \mathcal{V}$, $V \in \mathbf{DE}(V)$.)

10. The probability function Pr is just a joint distribution over the values of the variables in $\mathcal{V}$ which satisfies (MC).

    (a) As a matter of convention, it is standard to write something like

    $$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y)$$

    to mean that, for any values $x, y$, of $X$ and $Y$, the probability that both $X = x$ and $Y = y$ is just the probability that $X = x$ multiplied by the probability that $Y = y$. That is, the above means:

    $$\forall x, y \quad \Pr(X = x \wedge Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$$

    (b) What the (MC) tells us, then, is that, so long as $Y$ is not one of $X$'s causal descendants, for any value of $X$, $x$, any value of $Y$, $y$, and any value of $\mathbf{PA}(X)$, $\mathbf{p}$, the proposition $X = x$ is independent of the proposition $Y = y$ conditional on $\mathbf{PA}(X) = \mathbf{p}$.

    $$\forall x, y, \mathbf{p} \quad \Pr(X = x \mid Y = y \wedge \mathbf{PA}(X) = \mathbf{p}) = \Pr(X = x \mid \mathbf{PA}(X) = \mathbf{p})$$

    (c) Note that this probabilistic condition is rather different from the claims of the probabilistic theories of causation.

    i. Those theories attempted to go *from* information about probability (perhaps together with some, but not all, causal information) *to* complete information about the world's causal structure.

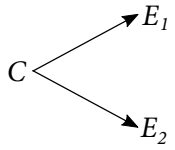
    ii. With the (MC), we provide the world's causal structure first (all of it), and then some probabilistic consequences follow. What the (MC) provides is not any kind of reduction of causal notions to probabilistic notions. Rather, the (MC) tells us something interesting about the way that probability and causation are interrelated.

    (d) Nevertheless, if the (MC) is true, some of the ideas from the literature on probabilistic theories of causation may be vindicated.

    i. We saw that Suppes (1970) (following Reichenbach (1956)) thought that, if $E_1$ and $E_2$ are joint effects of a common cause (that is, if the causal graph from figure 9.2 is correct), then, while $E_1$ and $E_2$ may be probabilistically dependent, they will be independent once we conditionalize upon $C$. And this is precisely what the (MC) says. For, the only causal parent of $E_1$ is $C$, so any probability distribution which satisfies (MC) with respect to the graph in figure 9.2 will be such that

    $$\Pr(E_1 \mid C, E_2) = \Pr(E \mid C)$$

**Figure 9.2** Joint Effects of a Common Cause




(e) There is an equivalent statement of the (MC) which will prove helpful. Iff the (MC) is true, then we may easily calculate the entire joint probability distribution over the variables in $\mathcal{V}$ by just using (what we can call) the 'transition probabilities' $\Pr(V \mid \mathbf{PA}(V))$.

i. That is: the following is equivalent to the (MC):

$$\Pr(V_1, V_2, \ldots, V_N) = \prod_{i=1}^{N} \Pr(V_i \mid \mathbf{PA}(V_i)) \qquad \text{(FT)}$$

('FT' stands for 'Factorization Theorem') The equivalence of (MC) and (FT) is proven in the appendix.

(f) What (FT) says is that, if we want to specify the entire joint probability distribution over the variables $V_1, V_2, \ldots, V_N$, we need only specify certain conditional probabilities—specifically, the probability of each variable taking on such-and-such value, given that its causal parents have taken on such-and-such values (for all possible substitutions of 'such and such').

(g) For instance, if we wish to calculate the joint distribution over $W, X, Y$, and $Z$, then:

i. if the directed edges are as shown in figure 9.1(a), then

$$\Pr(W, X, Y, Z) = \Pr(W) \cdot \Pr(X \mid W, Z) \cdot \Pr(Y \mid W) \cdot \Pr(Z \mid W)$$

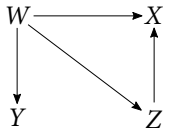
ii. and if the directed edges are as shown in figure 9.1(b), then

$$\Pr(W, X, Y, Z) = \Pr(W) \cdot \Pr(X \mid Y, Z) \cdot \Pr(Y) \cdot \Pr(Z \mid Y)$$

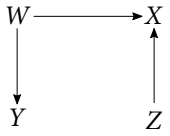
11. In causal Bayes nets, we can talk not only about *conditional* probabilities (which are defined in the usual way); but additionally about *intervention* probabilities.

(a) The way to think about an intervention probability is in precisely the way we previously thought about interventions in structural equations models—they *break the arrows* leading into a variable, severing its dependence upon its causal parents, and have their values set directly.

(b) To think through what this might entail, let's think about the causal structure from figure 9.1(a) (reproduced below)




(c) Suppose that we want to consider the intervention probability which results from setting $Z$ to $z$. That is: we want to take the distribution Pr, and calculate the distribution $\Pr_{Z=z}$. We should want a distribution in which $\Pr_{Z=z}(Z = z) = 1$, and in which this information only travels 'downstream' (so to speak). So we'll want a distribution which matches our original distribution as much as possible, but in which the probability that $Z = z$ is not affected by the probability that $W = w$, for any $w$. So we'll want a distribution which satisfies the (MC) (and therefore, the (FT)) over *this*, mutilated graph:




and in which $Z = z$ has probability 1.

(d) Applying the (FT) to the mutilated graph above, and stipulating that we want the same 'transition probabilities', we get that

$$\Pr{}_{Z=z}(W, X, Y, Z) = \Pr(W) \cdot \Pr(X \mid W, Z) \cdot \Pr(Y \mid W) \cdot \Pr{}_{Z=z}(Z)$$

since $\Pr_{Z=z}(Z = z) = 1$, this means that:

$$\Pr{}_{Z=z}(W, X, Y, Z) = \Pr(W) \cdot \Pr(X, W, Z) \cdot \Pr(Y \mid W)$$

(e) We'll use this general procedure to define what we mean by *intervention probability*. So, we'll say that, if we have a causal Bayes net $\mathcal{B}$, with probability distribution Pr, satisfying the (MC), we'll define

$$\Pr{}_{\mathbf{V}=\mathbf{v}}(V_1, V_2, \ldots, V_N) \stackrel{\text{def}}{=} \Pr{}_{\mathbf{V}=\mathbf{v}}(\mathbf{V} = \mathbf{v}) \cdot \prod_{V \notin \mathbf{V}} \Pr(V \mid \mathbf{PA}(V))$$

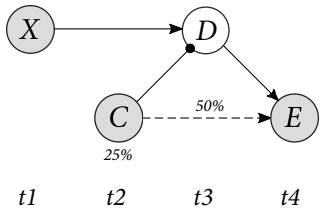
and we stipulate that $\Pr_{\mathbf{V}=\mathbf{v}}(\mathbf{V} = \mathbf{v}) = 1$.

## 9.3 Woodward's Interventionism

12. The formalism of causal Bayes nets allows us to interpret Menzies & Price's claim that $c$ caused $e$ iff $\Pr_c(e) > \Pr_{\neg c}(e)$ (at least, in the special case where there are binary variables corresponding to $c$ and $\neg c$—something would have to be said about how to generalize the account past binary variables).

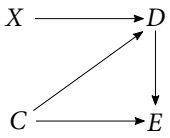
13. If we are thinking about the claim in terms of the singular causal relation, then there are good reasons to be skeptical of Menzies & Price's thesis—we've seen them already in our discussion of Lewis's 1986a theory of probabilistic causation.

**Figure 9.3** *C*'s firing caused *E*'s firing.




(a) Recall the probabilistic neuron diagram from that discussion, reproduced in figure 9.3.

(b) Such a neuron diagram may be modeled with the a causal Bayes net containing four binary variables, $X, C, D$, and $E$ (with the natural interpretation), the following set of directed edges:




and a probability distribution chracterized by the following 'transition probabilities' (here, I use '$V$' to stand for '$V = 1$' and '$\overline{V}$' to stand for '$V = 0$'):

$$\begin{aligned} \Pr(X) &= 1 & \Pr(C) &= 0.25 \\ \Pr(D \mid C, X) &= 0 & \Pr(E \mid C, D) &= 1 \\ \Pr(D \mid \overline{C}, X) &= 1 & \Pr(E \mid \overline{C}, D) &= 1 \\ \Pr(D \mid C, \overline{X}) &= 0 & \Pr(E \mid C, \overline{D}) &= 0.5 \\ \Pr(D \mid \overline{C}, \overline{X}) &= 0 & \Pr(E \mid \overline{C}, \overline{D}) &= 0 \end{aligned}$$

(c) Then, as is already clear from the diagram, it turns out that

$$\Pr_C(E) = 1/2 \quad \text{and} \quad \Pr_{\overline{C}}(E) = 1$$

(d) So the account tells us, falsely, that $C$ did not cause $E$.

(e) In fact, examples like this seem to undercut some of the primary motivation for the agency theory (understood as a theory about the singular causal relation). If you were an agent who had control over whether $C$ fired, and you cared only about $E$ firing, making $C$ fire would not be an effective means to your desired end. Nevertheless, $C$'s firing *did cause* $E$'s firing.

14. Even if Menzies & Price's account didn't run into these kinds of troubles, we might doubt that it is very illuminating once formulated in terms of causal Bayes nets. After all, the graphs from causal Bayes nets appears to already bake in a ton of causal

information—they bake in complete information about which variables causally determine the values of which other variables in the graph. Even if the account were true, it wouldn't provide us with an account of where the directed edges in our causal Bayes net come from.

15. Woodward (2003) provides a (non-reductive) interventionist account of these directed edges. That is, he provides an interventionist account of the relations of causal determination encoded in either a structural equations model or a causal Bayes net. That account is explicitly and intentionally non-reductive. However, while the account is circular, the circle is quite large—it winds its way through the notion of causal determination (the directed edges in our causal Bayes nets), singular, or token, causation, and the causal notion of an *intervention*.

16. To enter the circle at one point (any other would do, as well, since there is no pretense of reduction here): Woodward gives an account of singular, or token, causation which is identical to the second account we saw from Hitchcock (2001b). That is, he provides what we've called a *counterfactual counterfactual* account of singular causation; and, on this account, the acceptable assignments $\mathbf{g}$ to the off-path variables $\mathbf{G}$ are just those which do not make a difference to the on-path variables.

17. Next, Woodward characterizes the directed edges in a causal Bayes net, or a structural equations model—call this relation the relation of '*direct causal determination*'—as follows:

> Direct Causal Determination
> Relative to a variable set $\mathcal{V}$, $P$ directly causally determines the variable $Y$ iff there is some assignment of values $\mathbf{v}$ to all variables in $\mathcal{V}$ other than $P$ and $Y$ such that:
>
> (i) There is some intervention on $P$ (with respect to $Y$) which will change the probability distribution over $Y$ *even when all other variables in* $\mathcal{V}$ *are held fixed at the values* $\mathbf{v}$.

Woodward also defines another notion which will be important here. That is the idea of *contributing* causal determination. To understand this notion, think of Hesslow (1976)'s birth-control/pregnancy/thrombosis case. Changes to whether you have taken birth-control do not make a difference to your probability of thrombosis; though, in some sense, birth-control causes thrombosis. The sense in which birth-control causes thrombosis is that birth-control is a *contributing* causal determinant of thrombosis.

> Contributing Causal Determination
> Relative to a variable set $\mathcal{V}$, $A$ is a contributing causal determinant of $Y$ iff there is some direct-causal path from $A$ to $X$, $A \rightarrow V_1 \rightarrow V_2 \rightarrow \cdots \rightarrow V_N \rightarrow Y$, and some assignment of values $\mathbf{v}$ to all variables in $\mathcal{V}$ other than those on this path, such that:
>
> (i) There is some intervention on $A$ (with respect to $Y$) which will change the probability distribution over $Y$ *even when all the variables not on the path from* $A$ *to* $Y$ *are held fixed at the values* $\mathbf{v}$.

(a) Both of these definitions rely upon the notion of an *intervention*. This, too, is a causal notion, but it is one that Woodward hopes to give an account of.

(b) Up to this point in the course, we have modeled the *effects* of interventions in our models, but we have not been including anything in our model which corresponds to whether an intervention has taken place. Woodward wishes to characterize the notion of an intervention using these models, so he wants to include this kind of information within the models. He does so with the idea of an *intervention variable* (for $X$'s influence on $Y$).

(c) We are told that an intervention variable (for $X$'s' influence on $Y$), '$I_{X,Y}$', is a variable such that it can set $X$'s value in a way that it entirely independent of $X$'s other causal parents (except those that lie on the path from $I_{X,Y}$ to $X$), and that does not influence or correlate with the value of $Y$ *except via* its influence and correlation with $X$.

> Intervention Variable
> $I_{X,Y}$ is an intervention variable for $X$'s influence on $Y$ iff:
> (i) $I_{X,Y}$ is a contributing causal determinant of $X$;
> (ii) $I_{X,Y}$ has certain values such that, when $I_{X,Y}$ takes on those values, $X$'s probability distribution is no longer sensitive to the values of those causal parents which do not lie on the path from $I_{X,Y}$ to $X$.
> (iii) Any path of direct causal determination from $I_{X,Y}$ to $Y$ passes through $X$.
> (iv) $I_{X,Y}$ is probabilistically independent of all contributing causal determinants of $Y$ which do not lie on the path from $I_{X,Y}$ to $X$ to $Y$.

(d) Finally, using the notion of an intervention variable, we get the following account of what we mean by an *intervention* on $X$ (with respect to $Y$). An intervention on $X$ (with respect to $Y$) is just the value of an intervention variable (for $X$'s influence on $Y$) which singularly causes $X$'s value.

> Intervention
> $I_{X,Y}$ is an *intervention on* $X$ (with respect to $Y$) iff:
> i. $I_{X,Y}$ is an intervention variable for $X$'s influence on $Y$; and
> ii. $I_{X,Y} = i$ singularly caused $X = x$.

## 9.4 Objections to Interventionist Theories

18. Menzies & Price (1993) consider various objections to interventionist theories of causation. Their responses all depend upon an analogy between the interventionist theory of causation and the dispositionalist theory of color.

(a) A dispositionalist about color (like, *e.g.*, *redness*), contends that what it is for something to be colored red is for it to have a disposition to cause phenomenal redness in observers like us.

(b) Note that the dipositionalist about color does not say that, without observers like us, there would not be color. Just as salt would still be soluble even if there were no water, objects could still be red even if there were no observers. Dispositions do not depend upon the existence of their triggering conditions.

19. *Objection 1*: The Interventionist confuses the *epistemology* of causation with the *metaphysics* of causation. Interventions are a crucial part of how we come to acquire causal *knowledge*—but that does not mean that they have anything to do with *what causation is* in the world.

    (a) *Reply*: Similar complaints could be raised against the dispositional theory of color. However, it is clear that the dispositional theory of color does not confuse metaphysics with epistemology; it simply claims that color properties are ones that are individuated in part by their disposition to cause certain experiences in us. Similarly, the agency theory does not confuse metaphysics with epistemology; it simply claims that causal relations are ones that are individuated in part by its relation to our experience as agents.

20. *Objection 2*: The interventionist theory is circular. It explains causation in terms of what (suitably idealized, perhaps) agents could *bring about*. But the notion of *bringing about* is a causal one.

    (a) *Reply 1* (favored by (Woodward, 2003)): So it is. But not all circular accounts are trivially or viciously circular. Circular accounts can be genuinely illuminating and non-trivial in spite of their circularity, if they tell us something interesting about the ways that various concepts are related to one another. For instance:

        > Density
        > The rational number $x$ is greater than the rational number $y$ iff there is some rational number $z$ such that $z$ is greater than $y$ and $x$ is greater than $z$.

        This biconditional does not reduce the notion of 'greater than' to other notions; nevertheless, it does tell us something interesting and non-trivial about the rational numbers: namely, that they are *dense*—between any two, there is a third. (This isn't true of the integers, for instance.) Similarly, even though Woodward's interventionist theory is circular, it still tells us interesting things about the relationship between singular causation, causal determination, and interventions.

    (b) *Reply 2* (favored by (Menzies & Price, 1993)): An account may be *conceptually* reductive—in the sense that it takes you from old concepts on the right-hand-side to new ones on the left-hand-side—even when similar notions appear on both sides of the biconditional, so long as the notions that appear on the right-hand-side of the biconditional are known *by ostention* or *demonstration*.

        i. For instance, if you try to explain why an *itch* is to somebody who is unaware of the notion (that is, if you try to *give* somebody the concept of an *itch*), you might try something like the following:

ITCH
You have an *itch* iff you have something that causes you to feel itchy.

They may object to this account by pointing out that you invoked 'itchiness' notions on the right-hand-side. But, they contend, they don't know what it is to feel itchy—they don't have any itch-like concepts at all. Fair enough, you can respond, but then you can proceed to let an ant bite them. While they are attending to the way the ant bite feels, you can say:

ITCH (V2)
You have an *itch* iff you have something that causes you to feel *like that*.

21. *Objection 3*: Not all causal relationships are exploitable by agents like us. The moon causes the tides, but there is no experiment we could perform to remove the moon from the Earth's orbit.

    (a) *Reply 1*: Similar worries beset the dispositional theory of color. Consider Kripke's "killer yellow". In both cases, the solution is the same.

    (b) *Reply 2*: The kinds of interventions or manipulations the theory imagines need not be *in practice feasible* interventions or manipulations. Rather, what we are interested in are *in principle* interventions.

22. *Objection 4*: Agency theories are objectionably anthropocentric. They imply that, were there no agents, there would be no causation.

    (a) *Reply*: ...

Let the rank of a variable in a causal Bayes net be the largest number of directed edges leading from an exogenous variable to it. That is, if $U$ is an exogenous variable, and $V$ is an endogenous variable, then

$$\begin{aligned} rank(U) &= 0 \\ rank(V) &= 1 + \max\{rank(X) \mid X \in \mathbf{PA}(V)\} \end{aligned}$$

Throughout, to keep things neat, we'll use '$v_i$' to stand for the proposition $V_i = v_i$; '$\mathbf{pa}(x)$' to stand for the proposition $\mathbf{PA}(X) = \mathbf{pa}(x)$', and so on. That is: if a variable is lower-case, then we are assuming that it has a particular value; whereas, if it is uppercase, then we are assuming that its values may be bound by a universal quantifier (for instance, if we write '$\mathbf{PA}(V_i)$', then we allow that some of $V_i$'s parents may not be assigned any particular value).

**Lemma 1.** *Where $V_1, V_2, \ldots, V_N$ are any subset of variables from $\mathcal{V}$,*

$$\sum_{v_1, v_2, \ldots, v_N} \prod_{i=1}^{N} \Pr(v_i \mid \mathbf{PA}(V_i)) = 1$$

Note that this lemma does not rely upon (FT) or (MC).

*Proof.* Construct an ordering amongst the variables in $\mathcal{V}$ according to their rank as follows: if $rank(U) \neq rank(V)$, then $U \succ V$ iff $rank(U) > rank(V)$ and $U \prec V$ iff $rank(U) < rank(V)$. For those variables $U, V$ such that $rank(U) = rank(V)$, choose some arbitrary ordering: either $U \succ V$ or $U \prec V$; it won't make any difference.

Let '$X_i$' be the $i$th variable in this ordering you've constructed. Then,

$$\sum_{v_1, v_2, \ldots, v_N} \prod_{i=1}^{N} \Pr(v_i \mid \mathbf{PA}(V_i))$$

is:

$$\begin{aligned} &\sum_{x_1} \Pr(x_1) \cdot \left( \sum_{x_2} \Pr(x_2 \mid \mathbf{PA}(X_2)) \cdots \left( \sum_{x_{N-1}} \Pr(x_{N-1} \mid \mathbf{PA}(X_{N-1})) \cdot \left( \sum_{x_N} \Pr(x_N \mid \mathbf{PA}(X_N)) \right) \right) \cdots \right) \\ &= \sum_{x_1} \Pr(x_1) \cdot \left( \sum_{x_2} \Pr(x_2 \mid \mathbf{PA}(X_2)) \cdots \left( \sum_{x_{N-1}} \Pr(x_{N-1} \mid \mathbf{PA}(X_{N-1})) \cdot 1 \right) \cdots \right) \\ &\quad \vdots \\ &= \sum_{x_1} \Pr(x_1) \cdot \left( \sum_{x_2} \Pr(x_2 \mid \mathbf{PA}(X_2)) \cdot 1 \right) \\ &= \sum_{x_1} \Pr(x_1) \cdot 1 \\ &= 1 \end{aligned}$$

□

**Lemma 2.** *If* (FT) *holds, then, for any subset of variables* $\mathbf{X} = \{X_1, X_2, \ldots, X_N\} \subset \mathcal{V}$*, if* $\mathbf{AN}(X_i)$ *is the set of* $X_i$*'s* causal ancestors*, and* $\bigcup_i \mathbf{AN}(X_i) = \{A_1, A_2, \ldots, A_M\}$*, then*

$$\Pr(\mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} \Pr(X_i \mid \mathbf{PA}(X_i)) \cdot \prod_{j=1}^{M} \Pr(A_j \mid \mathbf{PA}(A_j))$$

*That is: given an ancestrally-closed variable set,*[5] *the distribution over the variables in that set factorizes into the 'transition probabilities'.*

*Proof.* Let $V_1, \ldots, V_K$ be all those variables in $\mathcal{V}$ *besides* $X_1, \ldots, X_N$ and $A_1, \ldots, A_M$. Then,

$$\begin{aligned}\Pr(\mathbf{X}, \mathbf{A}) &= \sum_{v_1, \ldots, v_K} \Pr(X_1, \ldots, X_N, A_1, \ldots, A_M, v_1, \ldots, v_K) \\ &= \sum_{v_1, \ldots, v_K} \prod_{i=1}^{N} \Pr(X_i \mid \mathbf{PA}(X_i)) \cdot \prod_{j=1}^{M} \Pr(A_j \mid \mathbf{PA}(A_j)) \cdot \prod_{k=1}^{K} \Pr(v_k \mid \mathbf{PA}(V_k))\end{aligned}$$

The second line follows from the factorization theorem (FT). Then, from the assumption that $X_1, \ldots, X_N, A_1, \ldots, A_M$ is ancestrally-closed, we have that none of $v_1, \ldots, v_K$ appear in any of the terms $\Pr(X_i \mid \mathbf{PA}(X_i))$, nor in any of the terms $\Pr(A_j \mid \mathbf{PA}(A_j))$. So these terms may be pulled out of the sums, and we get

$$\prod_{i=1}^{N} \Pr(X_i \mid \mathbf{PA}(X_i)) \cdot \prod_{j=1}^{M} \Pr(A_j \mid \mathbf{PA}(A_j)) \cdot \sum_{v_1, \ldots, v_K} \prod_{k=1}^{K} \Pr(v_k \mid \mathbf{PA}(V_k))$$

By **Lemma 1**, the remaining sum is equal to 1, and we get that

$$\Pr(\mathbf{X}, \mathbf{A}) = \prod_{i=1}^{N} \Pr(X_i \mid \mathbf{PA}(X_i)) \cdot \prod_{j=1}^{M} \Pr(A_j \mid \mathbf{PA}(A_j))$$

□

**Proposition 1.** *The probability function* Pr *in a causal Bayes net* $\mathcal{B} =< \mathcal{V}, \mathcal{E}, \Pr >$ *satisfies the* (FT) *iff it satisfies the* (MC).

*Proof.* Assume that Pr satisfies the (MC). Then, order the variables in $\mathcal{V}$ in the manner specified in the proof of **Lemma 1**, and subscript the variables $X_i$ so that

$$X_1 \prec X_2 \prec X_3 \prec \cdots \prec X_N$$

Then,

$$\begin{aligned}\Pr(X_1, \ldots, X_N) &= \frac{\Pr(X_1, \ldots, X_N)}{\Pr(X_1, \ldots, X_{N-1})} \cdot \frac{\Pr(X_1, \ldots, X_{N-1})}{\Pr(X_1, \ldots, X_{N-2})} \cdots \frac{\Pr(X_1, X_2)}{\Pr(X_1)} \cdot \Pr(X_1) \\ &= \Pr(X_N \mid X_{N-1}, \ldots, X_1) \cdot \Pr(X_{N-1} \mid X_{N-2}, \ldots, X_1) \cdots \Pr(X_2 \mid X_1) \cdot \Pr(X_1)\end{aligned}$$

By the construction of $\prec$, the right-hand-side conditioned-upon variables in each term above are not in the descendants of the left-hand-side variable. So, by the (MC), we may replace each product in the above with '$\Pr(X_i \mid \mathbf{PA}(X_i))$', and we get:

$$\Pr(X_1, \ldots, X_N) = \prod_{i=1}^{N} \Pr(X_i \mid \mathbf{PA}(X_i))$$

[5] By 'ancestrally closed', I mean that, for any variable in the set, all of its causal ancestors are in the set as well.

which is just the (FT). This establishes the right-to-left-hand direction of the biconditional.

Next, assume that Pr satisfies the (FT). We then show that it satisfies the (MC). Pick any variables $X, Y$ such that $Y \notin \mathbf{DE}(X)$. Just to keep things tidy, let '$\mathbf{P}$' denote $X$'s causal parents. Then, let $A_1, A_2, \ldots, A_N$ be the variables in $(\mathbf{AN}(X) \cup \mathbf{AN}(Y)) \setminus \mathbf{PA}(X)$.

Then,

$$\begin{aligned}\Pr(X \mid Y, \mathbf{P}) &= \frac{\Pr(X, Y, \mathbf{P})}{\Pr(Y, \mathbf{P})} \\ &= \frac{1}{\Pr(Y, \mathbf{P})} \cdot \sum_{a_1, \ldots, a_N} \Pr(X, Y, \mathbf{P}, a_1, \ldots, a_N)\end{aligned}$$

By **Lemma 2**, then, we have:

$$\begin{aligned}\Pr(X \mid Y, \mathbf{P}) &= \frac{1}{\Pr(Y, \mathbf{P})} \sum_{a_1, \ldots, a_N} \Pr(X \mid \mathbf{P}) \cdot \Pr(Y \mid \mathbf{PA}(Y)) \cdot \Pr(\mathbf{P} \mid \mathbf{PA}(\mathbf{P})) \cdot \prod_{i=1}^{N} \Pr(a_i \mid \mathbf{PA}(A_i)) \\ &= \frac{1}{\Pr(Y, \mathbf{P})} \cdot \Pr(X \mid \mathbf{P}) \cdot \sum_{a_1, \ldots, a_N} \Pr(Y \mid \mathbf{PA}(Y)) \cdot \Pr(\mathbf{P} \mid \mathbf{PA}(\mathbf{P})) \cdot \prod_{i=1}^{N} \Pr(a_i \mid \mathbf{PA}(A_i))\end{aligned}$$

And, by **Lemma 2** again, this is

$$\begin{aligned}\Pr(X \mid Y, \mathbf{P}) &= \frac{1}{\Pr(Y, \mathbf{P})} \cdot \Pr(X \mid \mathbf{P}) \cdot \sum_{a_1, \ldots, a_N} \Pr(Y, \mathbf{P}, a_1, \ldots, a_N) \\ &= \frac{1}{\Pr(Y, \mathbf{P})} \cdot \Pr(X \mid \mathbf{P}) \cdot \Pr(Y, \mathbf{P}) \\ &= \Pr(X \mid \mathbf{P})\end{aligned}$$

And so the (MC) holds. This establishes the left-to-right-hand direction of the biconditional. □

# 10 Causal Contrastivism

## 10.1 Context, Focal Stress, and Contrasts

1. We've already encountered cases in which context or contrast appear to make a difference to the truth-value of a causal claim (in our discussion of Mackie's notion of a *causal field*). To refresh our memories, let's look at a sampling of the data. We can illustrate with reference to the following case:

   > Susan the Bike Thief
   > There is a shop that sells bicycles and mopeds. After hours, there are two security systems: a security camera watches the bikes, while the mopeds are equipped with monitors which will trigger an alarm if they are removed from the store. Susan breaks into the store after hours and steals a bike. The guard sees her on the security camera, calls the police, and she is arrested.

2. Context-sensitivity:

   (a) Consider the following two monologues:

      i. "You know, they only had security cameras on the bicycles. So Susan could have stolen a moped, rather than a boke, and she would have gotten away scot free. It's s shame. Stealing the bike caused her to be arrested."

      ii. "You know, they have really good payment plans for bicycles. If she wanted a bike, then Susan could have easily afforded it. It's a shame. Stealing the bike caused her to be arrested."

   (b) It appears that the claim "Stealing the bike caused her to be arrested" is false when uttered in (2(a)i), yet true when uttered in (2(a)ii).[1]

   (c) Conclusion: context can affect the meaning of causal claims.

3. Focal stress:

   (a) Consider the following two claims:

      i. "Susan's stealing *the bike* caused her to be arrested."

      ii. "Susan's *stealing* the bike caused her to be arrested."

[1] *Cf.* Dretske (1977), Hitchcock (1996), Schaffer (2005), and Schaffer (2012).

(b) (3(a)i) appears to be false, whereas (3(a)ii) appears to be true.

(c) Conclusion: focal stress can affect the meaning of causal claims.

4. Contrastive 'rather than' clauses:

   (a) Consider the following two claims:

      i. Susan's stealing the bike, rather than the moped, caused her to be arrested.

      ii. Susan's stealing, rather than buying, the bike caused her to be arrested.

   (b) (4(a)i) appears to be false, whereas (4(a)ii) appears to be true.

   (c) Conclusion: the explicit listing of contrasts can make a difference to the meaning of otherwise identical causal claims.

5. Today, we're interested in how theories of causation can incorporate this data. We've already seen one approach—on Mackie (1965)'s theory, causal claims are made relative to a *causal field*, and changes in context or focal stress can make a difference with respect to the causal field. (Though Mackie didn't consider the case of focal stress or explicit contrast clauses, we saw that it can be naturally accommodated within his framework.)

## 10.2 Fine-Graining the Causal Relata

6. Dretske (1977) called attention to the fact that focal stress affects the meaning of causal claims. He was assuming that the meaning of nominalizations like "Susan's stealing the bike", and "her arrest" were just the entities they referred to.

   (a) Then, Dretske argues against the view that "Susan's *stealing* the bike" and "Susan's stealing *the bike*" refer to events. For, if they do, then they must refer to same event,

   $$[\![ \text{ Susan's } \textit{stealing} \text{ the bike } ]\!] = [\![ \text{ Susan's stealing } \textit{the bike} \text{ } ]\!]$$

   (b) Dretske additionally assumes the co-referring expressions can be substituted for one another without affecting the truth-value of the sentences in which they are embedded (because previous generations of philosophers loved Latin so much, we express this by saying that co-referring expressions are *intersubstitutable salva veritate*).

   (c) But, if so, then we'd have to say that (3(a)i) and (3(a)ii) are either both true or both false.

   (d) Dretske takes this to be a *reductio* of the assumption that "Susan's *stealing* the bike" and "Susan's stealing *the bike*" refer to events.

   (e) Rather, he contends that they refer to *event allomorphs*, and that event allomorphs are the causal relata.

(f) Hitchcock (1996) points out that this argument works equally well against the assumption that "Susan's *stealing* the bike" and "Susan's stealing *the bike*" refer to *facts* (at least, on Bennett (1988)'s theory of facts). For, at least on Bennett (1988)'s theory of facts, fact $f_1$ is identical to fact $f_2$ iff each proposition entails the other. And "Susan *stole* the bike" entails and is entailed by "Susan stole *the bike*".

   i. On other theories of facts, this will not be as clear. For instance, suppose that we treat propositions as functions from possible worlds to {0, 1}. And suppose that we think that, in a possible world in which a presupposition of a sentence is false, the proposition expressed by that sentence is neither true nor false (*i.e.*, the function doesn't map that possible world to anything). Then, consider the proposition "Susan *stole* the bike". This presupposes that Susan did something to the bike (see below). So, there is a possible world in which Susan stole the moped, but did nothing to the bike, and the proposition expressed by "Susan *stole* the bike" is neither true nor false at this world. However, the proposition expressed by "Susan stole *the bike*" is false at that world. So "Susan *stole* the bike" corresponds to a different function from possible worlds to {0, 1} than "Susan stole *the bike*" does. So these facts are not identical.

7. It's unclear how to use Dretske (1977)'s solution to diagnose the cases of context-sensitivity or contrast-sensitivity. Perhaps we could say that "Susan's stealing, rather than buying, the bike" refers to an event allomorph; and we could suggest that sometimes, context makes certain event allomorphs salient referents for otherwise unstressed nominalizations.

## 10.3 Increasing the Arity of the Causal Relation, part 1: Contrastive Causes

8. Dretske (1977) tried to capture the data from §10.1 by making the causal relata more fine-grained.

9. Another approach, championed by Hitchcock (1996) and Schaffer (2005, 2012), is to leave the causal relata alone but add additional *argument places* to the causal relation.

10. Hitchcock (1996) argues that we need an additional argument place for a contrast to the *cause*.

    Contrastive Cause
    Causal claims have the logical form "$c$, rather than any of $\mathbf{c}$, caused $e$" (where $\mathbf{c}$ is a set of potential *contrast events*).

11. Hitchcock's central motivation for Contrastive Cause derives from his desire to generalize the probabilistic theory of causation to non-binary variables.

(a) Recall, on probabilistic theories like those endorsed by Eells (1991), Cartwright (1979), and Skyrms (1980a), $C$ is a (positive) cause of $E$ iff, for some or all *causally homogenous background contexts* $K_i$,

$$\Pr_{K_i}(E \mid C) > \Pr_{K_i}(E \mid \neg C)$$

i. A causally homogenous background context $K_i$ specifies the value of all the causes of $E$ besides $C$.

ii. '$\Pr_{K_i}$' is just $\Pr(- \mid K_i)$.

(b) Hitchcock believes that this gets the wrong verdict in cases where the cause and effect variables may take on more than two values.[2] For instance:

i. Suppose that John can either smoke no packs a day ($S = 0$), one pack a day ($S = 1$), or two packs a day ($S = 2$). Then, the probabilistic theory says that smoking one pack a day is a positive cause of cancer $C = 1$ iff:

$$\Pr_{K_i}(C = 1 \mid S = 1) > \Pr_{K_i}(C = 1 \mid S = 0 \vee S = 2)$$

But whether this is so will depend upon how likely John was to smoke 0 packs a day or 2 packs a day. If John was very likely to smoke 2 packs a day, then his smoking one pack a day will be a negative causal factor for cancer; whereas, if John was very likely to smoke 0 packs a day, then his smoking 1 pack a day will be a positive causal factor for cancer.

ii. Hitchcock wishes to say that whether smoking one pack a day is a positive causal factor for cancer depends upon the relevant *contrast*. If the relevant contrast is smoking no packs a day, then it is, since

$$\Pr_{K_i}(C = 1 \mid S = 1) > \Pr_{K_i}(C = 1 \mid S = 0)$$

On the other hand, if the relevant contrast is smoking 2 packs a day, then it is not, since

$$\Pr_{K_i}(C = 1 \mid S = 1) < \Pr_{K_i}(C = 1 \mid S = 2)$$

(c) The basic contrastivist probabilistic theory of causation which Hitchcock advocates (to be fleshed out more later on) is the following:

> Contrastive Probabilistic Causation
> $C = c$, rather than $C = \mathbf{c}'$, where $\mathbf{c}'$ is a set of alternative values of $C$, caused $E = e$ iff
>
> $$(\forall c' \in \mathbf{c}')\, \Pr_{K_i}(E = e \mid C = c) > \Pr_{K_i}(E = e \mid C = c')$$

12. Hitchcock outlines a semantic theory of how focal stress introduces contrasts, and shows how this can be used to give a semantics for causal claims within a probabilistic theory of causation—one which explains the data from §10.1.

[2] Note: the above account was defended as an analysis of *type* causation—however, Hitchcock wishes to use it as an account of *token* causation here.

(a) We should distinguish the *content* of an utterance from its *presuppositions*. Consider:

> My sister got a divorce.

This sentence *says* that my sister got a divorce. However, it *presupposes* that I have a sister and that she was married. Why not say that the fact that I have a sister and that she was married is a part of the *content* of the utterance? Because, when you negate the sentence,

> My sister didn't get a divorce.

It *still* presupposes that I have a sister and that she was married. But, if the fact that I have a sister and that she was married were a part of the *content* of the utterance, then it would not be preserved under negation.

(b) On a theory of propositions which identifies them with functions from possible worlds to truth-value, $\{0, 1\}$, we can say that sentences with presuppositions are *partial* functions.

i. *E.g.*, $[\![$ My sister got a divorce $]\!]$ only maps worlds to $\{0, 1\}$ if they are worlds at which I have a sister and she was married.

Call the presupposition of an utterance $s$ '$\mathscr{P}(s)$'. Then, on this theory, $\mathscr{P}(s) = \{w \mid [\![s]\!]^w \in \{0, 1\}\}$.

(c) Just as propositions can have presuppositions, *variables* can have presuppositions as well.

i. A proposition is, after all, on the possible worlds interpretation, just a binary variable. More generally, a variable can take on multiple values but still be partial, in the sense that it doesn't map every possibility to some real number. Then, the presupposition of a variable is just all those worlds that the variable maps to some value or other.

$$\mathscr{P}(V) = \{w \mid V_w \in \mathbb{R}\}$$

(d) Hitchcock's theory is that, when a variable in cause- or effect-position has a presupposition, we must conditionalize upon that presupposition before making the probabilistic comparisons. So, in general, to evaluate whether $C = c$ caused $E = e$, we must consider a function like

$$\Pr_{K_i}(E \mid C \wedge \mathscr{P}(E) \wedge \mathscr{P}(C))$$

Though, in the case of the cause variable, the proposition $\mathscr{P}(C)$ is redundant, as it is entailed by $C = c$, for every $c$. So we can instead simply focus on

$$\Pr_{K_i}(E \mid C \wedge \mathscr{P}(E))$$

(e) Incorporating the contrastivism, it will be true that $C = c$, rather than $C = \mathbf{c}'$ (for some set of values of $C$, $\mathbf{c}$), caused $E = e$, just in case

$$(\forall c' \in \mathbf{c}) \Pr_{K_i}(E = e \mid C = c \wedge \mathscr{P}(E)) > \Pr_{K_i}(E = e \mid C = c' \wedge \mathscr{P}(E))$$

(f) A linguistic claim: focal stress serves to introduce the unstressed portion of the utterance as a presupposition.

i. "Susan *stole* the bike" presupposes that Susan $\phi$-ed the bike, for some specified range of $\phi$'s. The specified range of $\phi$'s which are presupposed are the relevant *contrasts*. And it *asserts* that $\phi =$ stole.

A. Notice that, while "Susan didn't steal the bike" *doesn't* seem to imply that Susan did anything at all *vis-a-vis* the bike, "Susan didn't *steal* the bike" *does* seem to imply that Susan did something to the bike; it just denies that the thing she did to it was steal it.

(g) In our causal claims, then, focal stress, in either the cause or the effect, serves to indicate which presuppositions we should conditionalize upon before making the probabilistic comparisons necessary to say whether the claim is true.

i. Consider the sentence $s =$"Susan's *stealing* the bike caused her to be arrested". This presupposes that Susan $\phi$-ed the bike, for some $\phi$. Hitchcock therefore claims that this will be true iff, for all relevant values of $\phi$ other than *stole*,

$$\Pr_{K_i}(\text{Arrest} \mid \text{ Susan stole the bike}) > \Pr_{K_i}(\text{Arrest} \mid \text{ Susan } \phi\text{-ed the bike})$$

(h) This account allows Hitchcock to handle cases of focal stress in the effect position (without postulating a contrast for the effect, as Schaffer (2005) does).

i. Consider the sentence $s =$"Susan's lack of scruples caused her to *steal* the bike". This also presupposes the Susan $\phi$-ed the bike, for some $\phi$. Call this presupposition '$\mathscr{P}(s)$'. Call the proposition that Susan stole the bike 'steal'. And let '$S$' be a variable describing Susan's scruples or lack thereof (with the actual value $S = 0$). Then, Hitchcock claims that this will be true iff, for all relevant values of $S$, $s$,

$$\Pr_{K_i}(\text{steal} \mid S = 0 \wedge \mathscr{P}(s)) > \Pr_{K_i}(\text{steal} \mid S = s \wedge \mathscr{P}(s))$$

This is plausibly true.

ii. Consider, on the other hand, the sentence $s' =$"Susan's lack of scruples caused her to steal *the bike*". This presupposes that Susan stole $x$, for some range of $x$'s. Call this presupposition $\mathscr{P}(s')$. Then, using the same notation as above, Hitchcock claims that this claim will be true iff, for all relevant values of $S$, $s$,

$$\Pr_{K_i}(\text{steal} \mid S = 0 \wedge \mathscr{P}(s')) > \Pr_{K_i}(\text{steal} \mid S = s \wedge \mathscr{P}(s'))$$

Given that Susan stole *something*, her lack of scruples does not make it any more likely that she would have stolen a bike rather than a moped. So this causal claim is false.

iii. To reiterate: notice that Hitchcock's theory only relies upon the contrasts introduced by focal stress in the *cause* position.

## 10.4 Increasing the Arity of the Causal Relation, part 2: Contrastive Causes and Effects

13. Schaffer (2005) argues that we should include contrasts in *both* the cause position *and* the effect position. That is: he argues for:

> Contrastive Causes and Effects
> Causal claims have the logical form "$c$, rather than any of $\mathbf{c}'$, cause $e$, rather than any of $\mathbf{e}'$" (where $\mathbf{c}'$ and $\mathbf{e}'$ are sets of potential contrast events).

14. Unlike Hitchcock (1996), Schaffer discusses contrastivism within the context of a counterfactual theory of causation. The account he puts forward (not as an analysis of causation, but rather as a test theory to display the virtues of contrastivism), is this:

> Counterfactual Contrastivism
> The occurrent event $c$, rather than any of $\{c'_1, c'_2, \dots\}$, caused the distinct occurrent event $e$, rather than any of $\{e'_1, e'_2, \dots\}$ iff:
> (a) each of $c'_1, c'_2, \dots$ are relevant alternatives to $c$;
> (b) each of $e'_1, e'_2, \dots$ are relevant alternatives to $e$;
> (c) for each $c'_i$, there is some $e'_j$ such that $O(c'_i) \boxdot\!\!\rightarrow O(e'_j)$; and
> (d) for each $e'_i$, there is some $c'_j$ such that $O(c'_j) \boxdot\!\!\rightarrow O(e'_i)$.

In the simple case where each set of contrasts is a singleton, this reduces to: $c$, rather than $c'$, caused $e$, rather than $e'$, iff $c'$ is a relevant alternative to $c$, $e'$ is a relevant alternative to $e$, and

$$O(c') \boxdot\!\!\rightarrow O(e')$$

15. His argument for Contrastive Causes and Effects is simply that it allows us to both explain the kind of data from §10.1, and additionally allows us to resolve several problems which a counterfactual theory of causation would otherwise face. We'll go through (some of) those problems and look at Schaffer's proposed contrastivist resolution.

### 10.4.1 Coarse-Grained, World-Bound Events

16. Recall that Quine (1950) and Davidson (1967) put forward a theory of events on which an event is just a region of spacetime at a world.

    (a) On this theory, events are incredibly *coarse-grained* and *world-bound*.[3]

[3] The world-boundedness of events is not apparent from Quine (1950) or Davidson (1967)'s claims; however, Schaffer (2005)'s conception of events includes the stipulation (p. 316) that they are worldbound.

(b) They are *coarse-grained* because superficially different descriptions of a region of spacetime nevertheless refer to the same event. The ball's rotating and the ball's heating up both refer to the same event; as do McEnroe's serving and McEnroe's serving awkwardly.

(c) They are *world-bound* because the event cannot occur at any worlds other than the one at which it does. A nearby possible world in which McEnroe serves confidently is not a world at which his actual serve occurs.

17. We saw that there was reason to think that this theory of events is not fit for a theory of causation—especially on a counterfactual theory.

(a) The coarseness of grain of the events is problematic because McEnroe's serve and his awkward serve appear to differ causally. It is true to say that McEnroe's tension caused him to serve awkwardly; but it is false to say that his tension caused him to serve.

(b) The world-boundedness makes the counterfactual $\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$ trivially true for all actually occurring $c$ and $e$. For, if $c$ actually occurs, then the most similar possible world at which it doesn't occur will be some world other than the actual one; and, since $e$ is world bound, this will be a world at which $e$ does not occur either. So every event caused every other, on the counterfactual theory.

18. Counterfactual Contrastivism allows us to resolve both of these problems for the Davidsonian theory of events.

(a) The adjective "awkwardly" in the sentence "McEnroe's tension caused him to serve awkwardly" does not merely serve to describe the effect event; it additionally suggests the relevant effect contrast to be an event where McEnroe serves, but not awkwardly. Then, since it's true that, had McEnroe not been tense, he wouldn't have served awkwardly, the causal claim comes out true.

(b) In contrast, the description of the effect in "McEnroe's tension caused him to serve" suggests the relevant effect contrast to be an event in which McEnroe doesn't serve. Then, since it's false that, had McEnroe not been tense, he wouldn't have served, the causal claim comes out false.

(c) Though the events themselves are world-bound, the counterfactuals we consider, according to Counterfactual Contrastivism, are not ones of the form $\neg O(c) \mathrel{\Box\!\!\rightarrow} \neg O(e)$. Rather, we consider counterfactuals of the form $O(c') \mathrel{\Box\!\!\rightarrow} O(e')$. And, if $c'$ and $e'$ are not occurrent events, these counterfactuals will not be trivially true.

### 10.4.2 Absence Causation (Causation by Omission)

19. On the one hand, our causal thought and talk is shot through with absence causation; on the other hand, absence causation can be metaphysically mysterious. In particular, Schaffer lists the following three worries with absence causation:

(a) If absences can be causes, then what are they? (A relation needs relata.) One thought is that the absence is just whatever event took place *instead* of the absent event. So, if Barry's failure to water the plant caused it to die, then it was whatever Barry did *instead* of watering the plant that caused it to die. Suppose that, instead, Barry sang a song. Then, we'd have to say that Barry's singing a song caused the plant to die. But that seems wrong.

(b) Absences lack *oomph*. Causes need *oomph*. So absences can't be causes.

(c) Some cases of absence causation appear to be infused with normative considerations (*cf.* McGrath (2005)). We say that Barry's failure to water the plant caused it to die, but not Carlos's failure to water the plant (though the plant's death counterfactually depends upon both).

20. Schaffer argues that Counterfactual Contrastivism answers two of these problems (it does not help out with normativity).

(a) Barry's failure to water the plant just is Barry's singing to song. Nevertheless, the sentence

> Barry's failure to water the plant caused it to die.

is true, while the sentence

> Barry's singing a song caused it to die.

is false. The reason is that the substitution of co-referring terms in this case leads to a shift in context. Referring to the event as "Barry's failure to water the plant" suggests the natural contrast event of Barry's *watering* the plant. And, had Barry watered the plant, it wouldn't have died. On the other hand, referring to the event as "Barry's singing a song" suggests the natural contrast event of Barry's *not* singing a song (question: what event is this? not an negative event, since there are no such things—why not the watering of the plant?). And, had Barry not sung a song, the plant still would have died (or so the solution supposes).

(b) Though there is no *oomph* between Barry's singing a song and the plant dying (let's suppose), there is *would be oomph* between Barry's watering the plant and its living. (This is essentially the line taken by Dowe (2000) in his discussion of absence causation.)

### 10.4.3 Transitivity

21. Given that causation is a four-place relation, you may suspect that causation cannot be transitive, since transitivity is a property of *binary* relations alone.

22. However, Schaffer (2005) argues that we may formulate a 4-place variant of transitivity (the idea is roughly to think of the four place relation as a binary relation between *pairs* of events and contrast sets, and then claim that *this* relation is transitive.) More carefully:

Transitivity (Four Place)
for any events $c, d, e$, and any sets of events $\mathbf{c}', \mathbf{d}', \mathbf{e}'$,

$$\textsc{cause}(c, \mathbf{c}', d, \mathbf{d}') \wedge \textsc{cause}(d, \mathbf{d}', e, \mathbf{e}') \Rightarrow \textsc{cause}(c, \mathbf{c}', e, \mathbf{e}')$$

23. Recall the counterexamples to transitivity we encountered in our discussion of Lewis's theory of causation:

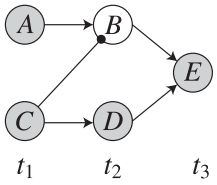
> Dog Bite
> A (right-handed) terrorist plans to detonate a bomb inside a building on Monday. On Sunday, a dog bites their right hand. So, on Tuesday, they detonate the bomb with their left hand. The building explodes. (McDermott 1995)
>
> Boulder
> A large boulder becomes dislodged, rolls down the hill, and careens towards a hiker. The boulder is large enough and fast enough that, if it hits the hiker, they will surely die. The hiker, seeing the boulder, ducks. The boulder flies over their head, and they survive unscathed. (attributed to an early draft of Hall 2004 by Hitchcock 2001b).
>
> Switch
> There is a switch which can be flipped to either the left or the right. Iff the switch is flipped to the left, then a lamp on the left will be turned on. Iff the switch is flipped to the right, then a lamp on the right will be turned on. Iff either lamp is on, then the room will be illuminated. The switch is flipped to the left, and the room is illuminated. (Pearl 2000)

(a) In Dog Bite, it appears that there is the following violation of transitivity:
   i. The dog bite caused terrorist to detonate the bomb with their left hand.
   ii. Terrorist's denotating the bomb with their left hand caused the building to explode.
   iii. The dog bite did not cause the building to explode.

(b) However, by paying attention to the contrasts in each claim, we can see that there is no violation of Transitivity (Four Place). For the first two causal claims are really of the following form:
   i. The dog biting, rather than not biting, caused the terrorist to detonate the bomb with their left hand, rather than detonating it with their right hand.
   ii. The terrorist detonating the bomb with their left hand, rather than not detonating the bomb, caused the building to explode rather than not.

(c) The account fares well with Dog Bite; however, it appears to fare less well with Boulder. In Boulder, there appears to be the following violation of Transitivity (Four Place):
   i. Boulder's falling, rather than remaining in place, caused Hiker to duck, rather than walking upright.
   ii. Hiker's ducking, rather than walking upright, caused Hiker to survive, rather than die.

**Figure 10.1** Preemption




iii. Boulder's falling, rather than not falling, did *not* cause Hiker to survive, rather than die.

(d) Schaffer's response is odd. He contends that the contrast event 'walking upright' refers to two different events in the first two causal claims.

i. In the first claim, the walking upright happens in a possibility where the boulder hasn't fallen.

ii. In the second claim, the walking upright happens in a possibility where the boulder has fallen.

iii. So, the two contrasts, in spite of being intrinsic duplicates, have different *relational* properties. So they are different.

(e) Let's just note that precisely the same thing happens in any case of causation without counterfactual dependence. For instance, in our well-worn case of preemption (figure 10.1), we might have hoped to invoke transitivity to say that *C*'s firing, rather than not firing, caused *E*'s firing rather than not firing. However, the following claims would not license that conclusion:

i. *C*'s firing, rather than not firing, caused *D*'s firing, rather than not firing.

ii. *D*'s firing, rather than not firing, caused *E*'s firing, rather than not firing.

for the contrast event in the first claim (*D*'s not firing) occurs in a world where *B* fires. While the contrast event in the second claim (*D*'s not firing) occurs in a world where *B* doesn't fire. In fact, its difficult to see what interesting theoretical work transitivity could ever do if Schaffer's response to Boulder were correct.

(f) Finally, it's worth noting that the response to Switch—which is offered for the case involving a train being diverted at a fork—works far less well in the kind of switching case involving the light bulb. It doesn't look like there is any untoward contrast-shifting going on in the following claims:

i. The switch's being set to the left, rather than the right, caused the left light to be on, rather than off.

ii. The left light's being on, rather than off, caused the room to be illuminated, rather than not illuminated.

iii. The switch's being set to the left, rather than the right, *did not* cause the room to be illuminated, rather than not illuminated.

(g) SCHAFFER could try to pull the same move here that he pulled with BOULDER, of course. But it's still worth noting that things look worse here than they did for the train case.

# Bibliography


Anscombe, G.E.M. 1969. "Causality and Extensionality." *The Journal of Philosophy*, vol. 66 (6): 152–159. [20]

Armstrong, David M. 1997. *A World of States of Affairs*, chap. 14: Singular Causation, 202–219. Cambridge Studies in Philosophy. Cambridge University Press, Cambridge. [4]

Bennett, Jonathan. 1988. *Events and their Names*. Hackett Publishers, Indianapolis. [3], [119]

—. 1996. "What Events Are." In *Events*, Roberto Casati & Achille C. Varzi, editors, 137–151. Dartmouth, Aldershot. [3]

Blanchard, Thomas & Jonathan Schaffer. forthcoming. "Cause without Default." In *Making a Difference*, Helen Beebee, Christopher Hitchcock & Huw Price, editors. Oxford University Press, Oxford. [78]

Briggs, Rachael. 2012. "Interventionist Counterfactuals." *Philosophical Studies*, vol. 160: 139–166. [75]

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies." *Noûs*, vol. 13 (4): 419–437. [120]

Collins, John, Ned Hall & L. A. Paul, editors. 2004. *Causation and Counterfactuals*. The MIT Press, Cambridge, ma. [130], [131], [132], [133]

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*. [19], [20], [123]

Dowe, Phil. 2000. *Physical Causation*. Cambridge University Press, Cambridge. [6], [8], [9], [43], [48], [49], [50], [51], [52], [53], [125]

Dretske, Fred I. 1977. "Referring to Events." *Midwest Studies in Philosophy*, vol. 2 (1): 90–99. [117], [118], [119]

Eells, Ellery. 1991. *Probabilistic Causality*. Cambridge University Press, Cambridge. [27], [35], [36], [37], [39], [120]

Elga, Adam. 2001. "Statistical Mechanics and the Asymmetry of Counterfactual Dependence." *Philosophy of Science*, vol. 68S: 313–24. [55]

Fair, David. 1979. "Causation and the Flow of Energy." *Erkenntnis*, vol. 14: 219–50. [8]

Gallow, J. Dmitri. 2015. "The Emergence of Causation." *The Journal of Philosophy*, vol. 112 (6): 281–308. [68]

—. 2016. "A Theory of Structural Determination." *Philosophical Studies*, vol. 173 (1): 159–186. [76]

Gibbard, Allan & William L. Harper. 1981. "Counterfactuals and Two Kinds of Expected Utility." In *Ifs: Conditionals, Belief, Decision, Chance, and Time*, William L. Harper, Robert C. Stalnaker & Glenn Pearce, editors, 153–190. Reidel, Dordrecht. [103], [104]

Hall, Ned. 2004. "Two Concepts of Causation." In Collins et al. (2004), 225–276. [60], [126]

—. 2007. "Structural Equations and Causation." *Philosophical Studies*, vol. 132 (1): 109–136. [76], [95], [100]

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. [100]

—. ms. "Appropriate Causal Models and Stability of Causation." [78]

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. [76], [79], [80], [82]

—. forthcoming. "Graded Causation and Defaults." *The British Journal for the Philosophy of Science*. [80]

Halpern, Joseph Y. & Judea Pearl. 2005. "Causes and Explanations: A Structural-Model Approach. Part 1: Causes." *The British Journal for the Philosophy of Science*, vol. 56: 843–887. [79], [80], [82], [83], [87], [93], [95], [100]

Hart, H.L.A. & Tony Honoré. 1985. *Causation in the Law*. Clarendon Press, Oxford, second edn. [14]

Hesslow, Germund. 1976. "Two Notes on the Probabilistic Approach to Causality." *Philosophy of Science*, vol. 43 (2): 290–292. [29], [35], [36], [110]

Hiddleston, Eric. 2005. "Causal Powers." *The British Journal for the Philosophy of Science*, vol. 56: 27–59. [91]

Hitchcock, Christopher. 1996. "The Mechanist and the Snail." *Philosophical Studies*, vol. 84 (1): 91–105. [117], [119], [120], [121], [122], [123]

—. 2001a. "A Tale of Two Effects." *Philosophical Review*, vol. 110 (3): 361–396. [3]

—. 2001b. "The Intransitivity of Causation Revealed in Equations and Graphs." *The Journal of Philosophy*, vol. 98 (6): 273–299. [60], [76], [79], [80], [82], [93], [95], [110], [126]

—. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review*, vol. 116 (4): 495–532. [100]

Hitchcock, Christopher & Joshua Knobe. 2009. "Cause and Norm." *Journal of Philosophy*, vol. 106 (11): 587–612. [98]

Huber, Franz. 2013. "Structural Equations and Beyond." *The Review of Symbolic Logic*, vol. 6 (4): 709–732. [75]

Hume, David. 1975. *A Treatise of Human Nature*. Clarendon Press, Oxford, second edn. [10]

Joyce, James M. 1999. *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge. [103]

Kim, Jaegwon. 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). [21]

Kitcher, Phillip. 1989. "Explanatory Unification and the Causal Structure of the World." In *Minnesota Studies in the Philosophy of Science*, Phillip Kitcher & Wesley Salmon, editors, vol. 13, 410–505. University of Minnesota Press, Minneapolis. [49]

Lewis, David K. 1973a. "Causation." *The Journal of Philosophy*, vol. 70 (17): 556–567. [14], [56], [57], [58], [59], [64], [66], [68], [80], [93], [95], [126]

—. 1973b. *Counterfactuals*. Blackwell Publishers, Malden, MA. [54], [55], [67], [75]

—. 1979. "Counterfactual Dependence and Time's Arrow." *Noûs*, vol. 13 (4): 455–476. [55], [58], [73]

—. 1981. "Causal Decision Theory." *Australasian Journal of Philosophy*, vol. 59 (1): 5–30. [103]

—. 1986a. "Causation." In *Philosophical Papers*, vol. II. Oxford University Press, New York. [16], [56], [58], [59], [60], [61], [62], [63], [66], [108]

—. 1986b. "Events." In *Philosophical Papers*, vol. II, 241–269. Oxford University Press, New York. [21], [56]

—. 2000. "Causation as Influence." *The Journal of Philosophy*, vol. 97 (4): 182–197. Reprinted in Collins et al. (2004, pp. 75–106). [66], [67], [92], [96]

—. 2004. "Causation as Influence." In Collins et al. (2004), chap. 3, 75–106. [56], [60], [62], [63], [64], [66], [67], [68], [69], [95]

Mackie, John L. 1965. "Causes and Conditions." *American Philosophical Quarterly*, vol. 2 (4): 245–55. [12], [16], [17], [80], [93], [95], [99], [100], [117], [118]

Maudlin, Tim. 2004. "Causation, Counterfactuals, and the Third Factor." In Collins et al. (2004), 419–443. [99]

—. 2007. "A Modest Proposal Concerning Laws, Counterfactuals, and Explanations." In *The Metaphysics within Physics*, 5–49. Oxford University Press, Oxford. [55]

McDermott, Michael. 1995. "Redundant Causation." *The British Journal for the Philosophy of Science*, vol. 46 (4): 523–544. [60], [126]

McGee, Vann. 1985. "A Counterexample to Modus Ponens." *The Journal of Philosophy*, vol. 82 (9): 462–471. [75]

McGrath, Sarah. 2005. "Causation by Omission: A Dilemma." *Philosophical Studies*, vol. 123: 125–148. [97], [98], [99], [100], [125]

Meek, Christopher & Clark Glymour. 1994. "Conditioning and Intervening." *The British Journal for the Philosophy of Science*, vol. 45: 1001–1021. [103], [104]

Mellor, D. H. 1995. *The Facts of Causation*. Routledge, London. [3], [19]

Menzies, Peter. 2004. "Causal Models, Token Causation, and Processes." *Philosophy of Science*, vol. 71 (5): 820–832. [79]

Menzies, Peter & Huw Price. 1993. "Causation as a Secondary Quality." *The British Journal for the Philosophy of Science*, vol. 44: 187–203. [102], [103], [104], [108], [109], [111], [112]

Mill, J. S. 1843. *A System of Logic*. [11]

Paul, L. A. & Ned Hall. 2013. *Causation: A User's Guide*. Oxford University Press, Oxford. [76]

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, second edn. [72], [88], [126]

Psillos, Stathis. 2009. "Regularity Theories." In *The Oxford Handbook of Causation*, Helen Beebee, Christopher Hitchcock & Peter Menzies, editors, chap. 7, 131–157. Oxford University Press, Oxford. [6]

Quine, W. V. 1950. "Identity, Ostension, and Hypostasis." *The Journal of Philosophy*, vol. 47 (22): 621–633. [123]

Quine, W.V.O. 1985. "Events and Reification." In *Action and Events: Perspectives on the Philosophy of Donald Davidson*, Ernie Lepore & Brian McLaughlin, editors, 162–171. Blackwell, Oxford. [20]

Reichenbach, Hans. 1956. *The Direction of Time*. Dover Publications, Mineola. [29], [43], [106]

Salmon, Wesley. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, Princeton. [43], [44], [45], [46], [47], [48], [49], [50]

—. 1994. "Causality without Counterfactuals." *Philosophy of Science*, vol. 61 (2): 297–312. [43], [50]

Schaffer, Jonathan. 2000. "Causation by Disconnection." *Philosophy of Science*, vol. 67 (2): 285–300. [51]

Schaffer, Jonathan. 2001. "Causation, Influence, and Effluence." *Analysis*, vol. 61 (1): 11–19. [68]

—. 2004. "Counterfactuals, Causal Independence and Conceptual Circularity." *Analysis*, vol. 64 (4): 299–309. [127]

—. 2005. "Contrastive Causation." *The Philosophical Review*, vol. 114 (3): 297–328. [117], [119], [122], [123], [125], [128]

—. 2012. "Causal Contextualism." In *Contrastivism in Philosophy*, Blaauw, editor, chap. 2, 35–63. Routledge. [117], [119]

Skyrms, Brian. 1980a. *Causal Necessity*. Yale University Press, New Haven. [120]

—. 1980b. "Higher Order Degrees of Belief." In *Prospects for Pragmatism*, D. H. Mellor, editor, chap. 6, 109–137. Cambridge University Press. [103]

Stalnaker, Robert C. 1968. "A Theory of Conditionals." In *Studies in Logical Theory*, N. Rescher, editor, chap. 4, 98–112. Oxford University Press, Oxford. [54], [75]

Strevens, Michael. 2003. "Against Lewis's New Theory of Causation: A Story with Three Morals." *Pacific Philosophical Quarterly*, vol. 84 (4): 398–412. [68]

—. 2008. *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, ma. [3]

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. North-Holland Publishing Company, Amsterdam. [27], [28], [29], [32], [33], [34], [106]

Varzi, Achille C. & Roberto Casati, editors. 1996. *Events*. Dartmouth, Aldershot. [131]

Weslake, Brad. forthcoming. "A Partial Theory of Actual Causation." *The British Journal for the Philosophy of Science*. [71], [80], [82]

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford. [104], [110], [111], [112]

Yablo, Stephen. 2002. "De Facto Dependence." *The Journal of Philosophy*, vol. 99 (3): 130–148. [76], [77], [92]

—. 2004. "Advertisement for a Sketch of an Outline of a Prototheory of Causation." In Collins et al. (2004), chap. 5, 119–138. [76], [77], [92]