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

References 92

Index 96
The Nomic Conditional and Natural Language Counterfactuals

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

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

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

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

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

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

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

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

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

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

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

\(^1\) Strictly speaking, similarity-based approaches can provide truth values for counterfactuals without a non-empty set of closest possible $C$-worlds. See discussion of the Limit Assumption in Lewis (69), p. 19–21.
The rest of this chapter is dedicated to clarifying how the nomic conditional is different from formalized versions of natural language counterfactuals vis-à-vis influence and causation. Along the way, I will suggest respects in which the nomic conditional is superior. Readers who are only interested in my positive account of causation or who find it highly implausible that the semantics of natural language conditionals could serve as a helpful constraint on one’s theoretical machinery for modeling influence may be able to skip ahead to the next chapter without too much loss.

### 11.1 Applicability to Singular Instances

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

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

### 11.2 Modus Ponens

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

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

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

11.3 Universal Modality

A well known feature of natural language counterfactuals is that they (at least sometimes) incorporate some structure akin to a universal quantifier. When Jill says, “If the glass had fallen, it would have broken,” one way that Jane can disagree is by responding, “No. What you said is incorrect. It might have survived.” Jill’s
counterfactual seems to mean something like, “If the glass had fallen, it definitely would have broken.” Incorporation of some structure that resembles a universal quantifier is common to most theories of counterfactuals. It is present in Goodman’s account, Lewis’s account, Gauker’s account, and others. All such theories hold that counterfactuals of the form, “If C were to occur, E would occur,” are true, roughly speaking, when all the relevant possible worlds where C holds, E also holds. For Goodman, the relevant possible C-worlds are the nomologically possible worlds with the appropriate (cotenable with C) background conditions. For Lewis, the relevant C-worlds are the C-worlds that are most similar to the actual world. For Gauker, the quantification ranges over linguistic contexts.

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

---

3 Lewis’s account advocates such a further restriction, so that some chance outcomes are counterfactually disallowed. It can turn out on his theory that in ordinary circumstances both of the following are true: “Had the vase been struck, it would not have broken,” and “Had the vase been struck, it would have been overwhelmingly likely to break.” Such results create a clear tension in the idea that counterfactuals incorporate an implicit ‘definitely’ in front of the consequent as noted above. More important, Lewis’s theory allows biasing the outcomes of fundamentally chancy processes to match the gross character of the actual material layout, and his only reason for including the bias in his system is to accommodate some naive intuitions like those concerning Morgenbesser’s coin, c.f., chapter 12.
future evolution is unusual. To rule them out just because of their unusual future behavior is hard to do without begging the question to some extent about what would have happened counterfactually.

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

### 11.4 Centering

For conditionals evaluated according to a truth-based semantics, the validity of modus ponens together with the validity of the inference rule, “From \( \alpha \& \beta \), infer \( \alpha \rightarrow \beta \),” implies that when the antecedent is true, the truth value of the counterfactual is equal to the truth value of the consequent. Applied to counterfactuals expressing influence, this means that if one starts out intending to evaluate the nomic consequences of a \( C \) that is actually instantiated, then (in order to obey truth-based modus ponens), the value of “If \( \widetilde{C} \) were to occur, then \( E \),” must come out equivalent to the truth value of \( E \). In Lewis’s (19) account of counterfactual logic, the semantic principle underlying this reasoning is labeled ‘centering,’ and it can be thought of in terms of comparative similarity as follows. There is one world that is most similar to the actual world: the actual world itself. One can extend the basic idea behind centering to alternative counterfactual logics as follows: Centering holds if and only if whenever \( \alpha \) holds, \( (\alpha \rightarrow \beta) \equiv \beta \) holds. (‘\( \equiv \)’ here represents the truth-functional biconditional.)

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

---

4 The most common objection is that the argument form, “From \( \alpha \& \beta \), infer \( \alpha \rightarrow \beta \),” allows one to take any two random truths and infer a counterfactual relation between them. Philosophers have a standard response, which is to argue that the inference is technically valid but that in natural language, such an inference carries an implicature that is not represented in the formal system. The seemingly invalid inferences turn out to be just those inferences that carry a misleading implicature. The inferences are always valid but can sometimes be misleading.
or non-existence of some event, so we only need to consider centering insofar as it concerns this more limited class of counterfactual conditionals.

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

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

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

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

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

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

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

\[ \text{Recall that I use the general term 'influence' to stand for the imprecise collection of all reasonable notions of influence, including our instinctive grasp of influence. 'Influence' is not a technical term in my system.} \]
11.5 Actuality-Focusing

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

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

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

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

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

Second, we only have epistemological access to a limited fraction of the coarse-grained states and virtually no access to fine-grained states. So if we are to learn about the causal regularities it will come by way of learning about influence insofar as it involves coarse-grained events. Furthermore, because strategies themselves are coarse-grained, (as discussed in §8.1), we need to be able to make assessments of counterfactuals involving coarse-grained events in order to assess the effectiveness of strategies.

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

### 11.6 Antecedents as Underspecified Propositions

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

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

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

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

Notice the following three points. First, there is nothing in the meaning of the antecedent, no matter how it is unpacked and clarified, that has anything to do with settling whether the bucket was dry in the counterfactual situation being entertained. If Jane had said afterward, "More precisely, I meant 'If I were to toss a ball into this bucket under circumstances where the janitor had previously filled the bucket with water, the ball would become wet,'" that would count as changing the topic. By all accounts, the dispute concerned the character of the bucket as it was when the statement was made. Second, the fact that the bucket was actually dry and was recognizable as such played an ineliminable role in the counterfactual's being demonstrably false. Third, any account of counterfactuals where Jane's claim does not come out false (or very improbable), and more generally fails to accord with most simple cases of counterfactuals dealing with interactions among physical objects, would be suspect because it would be unclear how that account would be relevant to influence. Although it is incorrect to dismiss an account of counterfactuals for not matching naive intuitions, in the particular subset of cases where the facts about influence are straightforward, a counterfactual intended to optimize talk of influence ought to make clear sense of them even if it does not render them explicitly true.
Setting theory aside, the example of the bucket shows that our natural reading of the counterfactual relies on facts concerning the actual layout of history in order to fill out the antecedent event. Thus, natural language counterfactuals employ underspecified antecedents. By contrast, no such deference to the actual material layout is built into my nomic conditional. There are instead just different antecedents one can consider. One can let \( C_w \) be a contextualized event instantiating Jane tossing the ball into a water-filled bucket and let \( \neg C_w \) be the contextualized event just like \( C_w \) except that Jane does not toss the ball. The contrastive event \( \tilde{C}_w \) is then just defined to be this ordered pair of contextualized events, whence the fundamental laws entail that the ball is much more likely to become wet. One is also free to consider the contrastive event \( \tilde{C}_d \), which is stipulated to be the ordered pair consisting of the contextualized event \( C_d \) where the bucket is dry with Jane tossing the ball in and the contextualized event \( \neg C_d \) that is identical except that Jane is not tossing the ball. The fundamental laws imply that \( \tilde{C}_d \) does not make the ball any more likely to become wet. One is also free to consider a contrastive event that represents a weighted mixture of the two possibilities, or a contrastive event like \( (C_w; \neg C_d) \). All these conditionals can be considered, with each representing a different relation of counterfactual dependence. None of them, though, are directed at representing the claim that Jane and Jill were disputing. In order to represent that claim, one would need to characterize the contrastive event not by specifying its condition directly in terms of various fine-grained events, but indirectly by specifying its elements as modifications to whatever state of the world exists at the time Jill and Jane are debating. Within the context of my account, one can make sense of their debate as follows. Jane is saying in effect that the actual state at \( t \) is such that if it were naturally coarse-grained to include her throwing the ball into the bucket, the ball would probably become wet. Looking in the bucket after \( t \) provides evidence that the actual state at \( t \) had a dry bucket, which would be contextualized into a dry-bucket-instantiating event like \( C_d \), which fixes a low probability of the ball becoming wet.

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

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

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

Accounts that only use sufficiently specified antecedents can make some sense of the implicit indexical reference to actuality in natural language by conceiving of those counterfactuals as incorporating a map in the sense employed by computer programmers. The kind of map relevant to counterfactual evaluation is an index of all the possible states at time $t$, linked to a state at $t$ that has been modified appropriately to instantiate the antecedent event. To evaluate ‘Jane tosses a ball into the kind of bucket her intention selects while everything else going on is like it is in

---

6 There is no requirement that a theory actually use all the resources listed in the table, only that they are in general free to use them.
the actual world at time \( t \), one would try to reckon, as best one can, the actual state of the world at \( t \), use the map to figure out what the modified state is, and then let the laws operate on that modified state. The map structure in effect allows us to take the indexical reference to actuality out of the semantics of the counterfactual and instead impose it as an external parameter that fixes which antecedent event is relevant given the actual material layout. Call this the ‘map approach.’

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

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

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

- Accounts that hold counterfactually fixed some actual facts after \( t \) or make the counterfactual evolution more likely to match actual facts than the laws indicate\(^\text{[5]}\)
- Accounts that use special science laws, folk psychological principles, or other similar rules of thumb for generating the counterfactual historical development
- Accounts that also evolve the modified actual state into the past
- Accounts that permit counterfactual backtracking, e.g. counterfactual inferences are made toward the past and then toward the future.
- Forking accounts
- Extended forking accounts

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

\[ \text{\footnote{ Even though some actual post-} t \text{ facts are carried over to the counterfactual history, the rest of the evolution should come from the laws.}} \]
The Nomic Conditional and Natural Language Counterfactuals

the antecedent obtaining at \( t \), and then later lawfully to the rest of the counterfac-
tual history. An extended forking account generalizes the applicability of forking
accounts to deterministic settings by permitting miracles to generate the fork.

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

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

The central problem to be solved by any account of counterfactuals that em-
ployers underspecified antecedents and purports to be optimized for understand-
ing influence and causation is how to get from the underspecified antecedent to
something approaching the results of the minimal account. There are two ways of
approaching this problem. The first, called ‘the informal approach,’ involves just
relying on our implicit ability to fill out the underspecified antecedent as needed
in order to match what intuitively seems like the correct (fleshed out) antecedent
event. It is not always transparently clear whose theories are intended to accord with which approach, but I think it is fair to associate the informal approach with a wide variety of theories, e.g. \[5\] \[8\] \[16\] \[48\].

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

Another principled theory is Lewis’s \[42\] theory of world-similarity.

In the rest of this section, I intend to demonstrate that both approaches are suboptimal for understanding causation. If correct, that will favor my account, which is an alternative to what these two models of counterfactuals share: the use of underspecified antecedents. In order to address both approaches, it is helpful to have an example that highlights where they diverge from each other. Suppose a bomb capable of very quickly destroying Earth is set to activate when a neutrino interacts with the trigger. Neutrinos are ubiquitous in nature but each only rarely interacts with ordinary matter. In actuality, the bomb never activates even though all it takes for activation is for a single neutrino to be ever so slightly in a different location. Now, consider an ordinary counterfactual having nothing to do with the bomb. Let us say that Guy completed an ordinary workday at the office with his ever-observant boss. By ordinary standards, the following is true:

> If Guy had not shown up for work, his boss would have noticed.

According to the informal approach, we are free to evaluate this counterfactual by examining situations where Guy spends the day goofing off or is sick or sets out for a new life of adventure or dies or some disjunction of these and other plausible ways Guy could fail to show up for work. We are not required to. For this particular example, most reasonable ways of filling out the antecedent state lead to the counterfactual being reasonably highly probable, or loosely speaking, true. (If one imposes the universal modality principle, it will

\[8\] If there is any fundamental chanciness, one conditionalizes on the truth of the antecedent to get the appropriate probability measure over future evolutions.
turn out false, because the boss would not necessarily notice Guy’s absence, but on such a reading, virtually all mundane counterfactuals turn out false.) By the principled approach, one needs to consult one’s theory of counterfactuals. If that theory identifies the bombed-Earth worlds as the theoretically specified worlds relevant to the semantic value of the counterfactual, then the counterfactual will be definitively false (or highly improbable) because the boss will not exist. This illustrates that the two approaches can differ on how to evaluate the same counterfactual statement.

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

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

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

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

The other known example of the principled approach is Lewis’s theory of world-similarity. The two core ideas behind Lewis’s theory of counterfactuals is that the semantics governing counterfactuals is based on truth conditions incorporating a relation of comparative similarity and that a principled theory of world-similarity can instruct us of the truth of counterfactuals pertaining to nomic interactions among worldly stuff. The semantics itself is simple enough. The antecedent and consequent are modeled as propositions. A counterfactual $C \rightarrow E$ is true if and only if $C$ is a contradiction or for some $(C&E)$-world, there are no possible $(C&\neg E)$-worlds that are more similar to actuality than it. $E$
counterfactually depends on \( C \) if and only if \( (C \implies E) \land (\neg C \implies \neg E) \). Even though the criticisms in the previous subsections suffice to show that Lewis’s logic and semantics are suboptimal for understanding influence in an empirical analysis, Lewis’s substantive theory of how to address underspecified antecedents is still worth some investigation. I will not address any of the alleged counterexamples to Lewis’s theory of world-similarity but will merely point out a single problem with his approach that will likely extend to any other principled approach to the problem of underspecified antecedents. The failure of both the informal and principled approach then casts doubt on the utility of any theory that employs underspecified antecedents, including all approaches to counterfactuals based on a similarity relation.

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

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

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

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

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

To summarize, there are two methods of handling antecedents as underspecified propositions, both of which fail. The principled approach suffers from an inability to pick out the relevant worlds and the informal approach only succeeds to the extent that it cheats by treating antecedents as sufficiently specified. Thus, treating antecedents as underspecified propositions is inferior to treating them as sufficiently specified. Because the nomic conditional treats antecedents as sufficiently specified, it is better suited for an empirical analysis of causation focused on effective strategies.
In this chapter, I will further illustrate the character of the nomic conditional using the example of Morgenbesser’s coin to highlight the most important differences between the nomic conditional and natural-language-based models of counterfactuals. The goal of this chapter is to inoculate the account of general causation presented in chapter 5 from some misinterpretations that can occur if readers are not fully alert to the significant respects in which prob-dependence differs from versions of counterfactual dependence based on standard models of counterfactual conditionals. In the chapter on causal asymmetry, I contrast the nomic conditional with Barry Loewer’s SM-conditional based on statistical-mechanical probabilities.

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

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

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

1. **Change Nothing:** The simplest method of evolving S forward in time is just
to fill in the future of S with exactly what happens in the actual world. This
method is virtually never employed when the antecedent is false—for obvi-
ous reasons—but when the antecedent is true, it is standard practice to let
S be the actual fine-grained state at t, even when the actual state does not
instatiate something that ought to count as among the situations contextu-
ally relevant to the evaluation of the counterfactual. Then, instead of using
the laws to evolve S forward, we just fill out its future with the actual future.
The ‘change nothing’ procedure has the consequence that when C is true,
the truth value of “If C had happened, E would have happened” matches
the truth value of E. This makes sense of several theorems of counterfactual
logic, especially \((C \land E) \vdash (C > E)\) and \((C \land \neg E) \vdash \neg (C > E)\), though I am
not taking sides on whether these are good principles of reasoning.

2. **Change Everything:** Another simple method for evolving S forward in time
is just to let the future be whatever S fixes for the future. This method corre-
ponds with the following way of reasoning about contrary-to-fact possibil-
ities: One imagines hypothetically going back to some time, fiddling with
the state if necessary to make the antecedent true in some contextually rele-
vant way, and then let the laws dictate what happens afterward without any
bias due to (later) accidental contingencies in the actual world. One is fig-
uratively rerolling the dice of nature for the counterfactual history starting
with S, as depicted in Fig. 12.1. Following this method, we would say that if
Jill had bet tails, there would have been a fifty percent chance of her winning,
but no fact of the matter as to which outcome would have happened.

---

9 Philosophers tend to emphasize that natural language counterfactuals have a universal-
quantifier character. When the ‘change everything’ procedure leads to no definite outcome for
Jane, we say it is false that she would have won and false that she would have lost. The implicit
3. **Bifurcated**: The bifurcated method is simply to apply the ‘Change Nothing’ method if the antecedent (construed as a proposition) is true, and to apply the ‘Change Everything’ method if it is false. This is motivated by the observation that the ‘Change Nothing’ method seems to match our intuitions well when the antecedent is true, but greatly mismatches our intuitions when the antecedent is false.

4. **Change Infected Regions**: It is possible to match people’s instinctive judgments of particular counterfactuals better than the previous options by evolving the antecedent state forward in time while selectively employing either the actual world as a guide or the dynamical laws. We can do so by drawing a distinction between an infected and uninfected region of spacetime. The way it works is as follows. We start with the actual state, $S@$, at time $t$ and check to see whether the actual world around $S@$ instantiates $C$. If it does, we declare our initial hypothetical state $S$ to be $S@$ and we declare that it counts as entirely uninfected. If $S@$ does not instantiate $C$, we set $S$ to be just like $S@$ except that we modify it to make it instantiate $C$ in some contextually appropriate way. One can typically accomplish the alteration by just modifying a local patch of physics and leaving everything else exactly the same, or one can back up in time and let the laws evolve to $t$ and then conditionalize on $C$. Whatever patch is modified to instantiate $C$ counts as infected, and everything else counts as uninfected. Now, we evolve $S$ forward in time to create a counterfactual scenario, a family of counterfactual worlds that count as the worlds relevant to the truth of $C > E$. The two rules for evolving the state forward in time are these: Wherever $S$ is uninfected, we just copy whatever happened in the actual world into the counterfactual scenario. Wherever $S$ is infected, we use the actual dynamical laws to

---

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

---

I am lifting this distinction from Maudlin, p. 30.
fill in the counterfactual scenario by evolving the material contents of the counterfactual scenario just prior to the occurrence of the infected region and propagate them lawfully into the infected regions. There are different prescriptions for the rules as to when a region of space-time counts as being infected. Let us now examine some possible rules for infection in order to get a better feel for the methodology. No matter which rules we use for how infection spreads, the resulting system satisfies the three previously mentioned principles common to counterfactuals based on natural language: centering, actuality-focusing, and antecedents as underspecified propositions.

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

- Infection as Probability-Affecting: A more restricted conception of infection (\textsuperscript{57} (\textsuperscript{38}) (\textsuperscript{50}) (\textsuperscript{51}) is that one should count as infected any event whose occurrence is probabilistically dependent on events from

\textsuperscript{11} Remember that non-spatiality is the principle that events never termine events that are at space-like separation.
infected areas. Because the chance of the coin landing tails is arguably very nearly the same whether Jane bets heads or tails, it is probabilistically independent of the bet and so is deemed uninfected. Because Jane's winning the bet does probabilistically depend on her bet when one holds fixed the actual flip outcome, it counts as infected. Thus, by using the laws to fill in the infected region, we get the result that the counterfactual comes out true. There are serious questions about how to make the notion of probabilistic dependence suitably precise without employing counterfactuals that are circular or relying on parameters that would leave the semantic value of the counterfactual too subjective, but I will leave these issues aside.

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

12 Which option is appropriate depends on whether one construes the counterfactual as having a universal-quantifier character as discussed in the footnote at the end of the ‘Change Everything’ option above.
It is frequently unclear whether the models listed above are intended by their advocates as psychological models for explaining people’s intuitive judgments about the correctness of counterfactual claims, or whether they are intended as part of a metaphysical model with consequences for influence and causation. Without a clear enough guide to their purpose, there is no way to decide which methods for assessing the Morgenbesser counterfactual are better than others.

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

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

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

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

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

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

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

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

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

---

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

14 In making this claim, I am not assuming that the fundamental laws disallow that what happens in one location causally contributes to what happens elsewhere. Quantum mechanical entanglement, for example, might ensure that everything in the universe is linked in such a way that a counterfactual alteration to one location determines differences everywhere else, but the scenarios being entertained in my discussion ex hypothesi do not fundamentally interact with one another in any interesting way.
the method of ‘infection as culpable causation’ both attempt to hold certain aspects of the actual future fixed under the counterfactual alteration.) In order to get the Morgenbesser counterfactual to come out true, the additional factors—no matter what else they include—must contain the actual outcomes resulting from fundamentally chancy processes. But it is exactly these actual outcomes that we do not want to include in our predictions of what happens in regions other than R because it is simply the nature of distinct fundamental chance outcomes (that are independent of one another insofar as fundamental causation is concerned) that what happens by chance in one case does not have any bearing on what happens by chance in other cases. So, the machinery needed to make the Morgenbesser counterfactual true works against our having an optimal guide for predicting and explaining what happens elsewhere.

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

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

15 Again, I am setting aside empirical phenomena relevant to the evaluation of our theories of how people think about counterfactuals.
We could overcome these previously discussed problems by revising the methods listed above to be far more liberal about what counts as a minimal modification to instantiate the antecedent. Specifically, let us now replace the step in the above methods that said, ‘alter the actual state minimally to make the proposition \( C \) true’ to ‘instantiate \( \overline{C} \) in place of the actual state’. When we do, we obviate problems with centering, actuality-focusing and antecedents as underspecified propositions.

With this modification in place, it turns out not to matter too much whether we employ ‘change everything’ or ‘infection by contribution’ in my method for evaluating prob-dependence. So long as \( \hat{C} \equiv (\overline{C}, \overline{\overline{C}}) \) occupies a portion of a single time slice, we will only get prob-dependence of events on \( \hat{C} \) when they are in the domain of influence of the region where \( \overline{C} \) and \( \overline{\overline{C}} \) disagree about the material facts. For example, suppose we have a naturally contextualized event, \( \overline{C} \), involving a match being struck in region \( R \) and a contrast \( \overline{\overline{C}} \) that is exactly the same as \( \overline{C} \) except for having a prototypical lack of a match strike in region \( R \). Furthermore, suppose the laws are relativistic and consider some consequent \( E \) that is a coarse-graining of a fundamentally chancy event \( e \) that actually occurred outside the light cone of \( R \) and to the future of \( \hat{C} \). According to my method, we evaluate both counterfactuals, \( \overline{C} \Rightarrow E \) and \( \overline{\overline{C}} \Rightarrow E \), by rerolling the dice of nature throughout the future of \( \overline{C} \) and \( \overline{\overline{C}} \), ignoring that \( E \) actually occurred. Although both counterfactuals will have a non-trivial value (equal to the chance given either \( \overline{C} \) or \( \overline{\overline{C}} \)), their values will be exactly the same, which implies no prob-dependence.\footnote{Recall the definition of a domain of influence from §2.6.}

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

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

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

---

\footnote{Recall the definition of a domain of influence from §2.6.}
\footnote{This equality holds because \( e \)’s domain of contributors, \( e \)’s past light cone, intersects with both \( \overline{C} \) and \( \overline{\overline{C}} \) in the same region including exactly the same material content. (Recall the definition of a ‘domain of contributors’ from §2.6.)}
The methods ‘Infection as Probability-Affecting’ and ‘Infection as Culpable Cause’ do not fare well even with the helpful modification in place. What is most problematic about them is that they provide insufficient precision. There is no guide as to how much probability-affecting counts as enough to trigger infection and because any threshold will be arbitrary, some clarificatory stipulation will need to be added. The same goes for clarifying the precise extent of the supposed relations of causal dependence governed by intuitions about culpability. With any such threshold parameters, there will be cases where a large difference in the value of the counterfactual will occur because of a small change in the standards for what triggers the spread of infection. This results in ungraceful conceptual degradation. Furthermore, the parameters to make either method precise have been notoriously hard to pin down. I think all such avenues of exploration amount to attempts to explain the clear in terms of the murky.

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

---

18 See §1.1.
the counterfactual dependence only arose because $E$ was unlikely, it is helpful to have a notion of counterfactual dependence that mitigates the undesired implication. By carrying over from the actual world to the counterfactual worlds any chance outcomes that would result in misleading counterfactual dependence, one automatically eliminates much of the spurious dependence. But because we need to allow that some chance outcomes are not held counterfactually fixed, we need a principle that rules out spurious dependence while also ruling in genuine dependence. Accomplishing that task, I take it, is the purpose of 'Infection as Probability-Affecting' and 'Infection as Culpable Cause.' Because such construals of counterfactual dependence are psychological kludges, it is not surprising that 'Infection as Probability-Affecting' and 'Infection as Culpable Cause' are unimpressive at clarifying the nature of causation and influence.
Orthodox Conceptual Analysis

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

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

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

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

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

I will now discuss two dimensions along which old-fashioned conceptual analysis can be relaxed. The first involves dropping the requirement that the analysis provide an explicit definition. Old-fashioned orthodox analyses have room to accommodate the vagueness of the target concept by way of the vagueness of the concepts in terms of which the explicit definition is formulated. But Frank Jackson (53) alone and with David Chalmers (7), for example, advocate permitting analyses that are vague in the degree of fit they make with the structure of necessary and sufficient conditions. This lowering of the bar is motivated by the recognition that our cognitive grip on some facts comes through a pattern recognition capacity rather than a rule checking capacity. For even the most mundane concepts, like our concept of the alphabetic character G, it is very difficult to write out an explicit definition of G, i.e. a specification of which glyphs count as clear cut instances of G holding across a wide variety of typefaces. Despite the seeming lack of an explicit rule for G-ness, people have widely shared opinions about the extension of G, with some vagueness at the borderline. This suggests that our concept of G exists by virtue of a shared capacity to recognize exemplars of G and to accommodate for variations from the exemplars. So, the more relaxed version of orthodox analysis Jackson defends allows the advocate of some analysis to incorporate fudge factors in the necessary and sufficient conditions that implicitly rely on our shared pattern-matching capacities. There is a danger that by permitting this kind of
latitude, the conditions of adequacy will not be principled enough to differentiate between informative analyses and those that are mere data fitting, but this by itself is not a problem for conceptual analysis. We have good enough pattern-matching capacities that allow us to distinguish between trivial and informative analyses, even given the flexibility provided by this relaxation. So, I am in full agreement with Jackson on the advisability of this relaxation of the standards for an adequate conceptual analysis.

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

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

It is clear enough—at least for present purposes—why someone interested in providing a conceptual analysis of our ordinary notion of causation should attend carefully to intuitions about cases. What we wish to emphasize is that even someone interested in “synthesizing” a new
and potentially useful concept needs to heed these intuitions, else she risks cutting her project free of any firm mooring. More specifically, a reasonable and cautious approach for her to take is to treat intuitions about cases as providing a guide to where interesting causal concepts might be found. Thus, although the account can selectively diverge from these intuitions, provided there are principled reasons for doing so, it should not diverge from them wholesale. (p. 31)

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

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

Although the quoted characterization of conceptual analysis from Collins, Hall, and Paul is compatible with empirical analysis, I believe an examination of the practices of orthodox conceptual analysts supports the conjecture that they are engaged in an activity significantly different from empirical analysis. Except for some quibbles, everything I have said so far about empirical analysis is also compatible with Jackson’s clarification (pp. 30–36) of conceptual analysis. Nevertheless, an examination of Jackson’s practical applications of his version of conceptual analysis, e.g. his discussion of color, makes clear that he is interested in locating a concept of color that fits folk intuitions about color much more closely than the chemist’s concept of metal oxides fits folk intuitions about rust. The same holds for discussions of causation by Collins (8), Paul (53), and earlier work of Hall, e.g. (23)(21), though Hall’s paper, “Two Concepts of Causation,” departs from the requirement that an adequate
conceptual analysis needs to systematize all the platitudes regarding causation as a single regimented concept.

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

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

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

13.1 The Orthodox Metaphysics of Causation

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

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

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

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

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

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

Consider the example known as Morgenbesser’s coin (53) p. 26 fn. 33. Suppose the world is governed by fundamentally chancy laws and that there is enough randomness in the microscopic world so that the fifty percent chance that an ordinary coin flip lands heads is overwhelmingly due to fundamental chanciness, not to our ignorance of the microscopic details of the setup. A coin is flipped and when it is airborne, Jane bets heads, and the coin lands tails. When people are told of such stories they tend to agree with the counterfactual, “If Jane had bet tails, she would have won.” Orthodox theories, e.g., (51) (59), tend to take such folk opinions as truths that need to be entailed by any adequate analysis, not merely as practices that are understandable as folksy approximations of some deeper structure that is relatively far removed from the explicit content of the counterfactual claim. Again, not all folk intuitions are sacrosanct according to new-fangled orthodox analysis, but to the extent causation is interpreted as part of a larger theory that includes counterfactual claims, the orthodoxy tries to impose some burden of explanation on theories that deny the Morgenbesser counterfactual.
13.1.2 CULPABLE CAUSES

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

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

19 The term ‘culpable cause’ has been used previously by Mark Alicke (14) to designate something altogether different: the psychological effect of perceived moral blameworthiness on judgments of causal impact.
not equivalent. When Guy won the lottery, his purchase of the lottery ticket was one of the causes of his winning but it was not an event that caused him to win. ‘Culpable cause’ refers to the ‘one of the causes’ disambiguation.

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

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

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

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

---

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

21 Dowe is clear in his Ch. 6 that intuitions about omissions can be satisfactorily vindicated without his account needing to make it explicitly true that the causation by omission is real causation.
be inadequate if it could not make sense of its utility. While I do not have the space to present an adequate comparison of my account with causal modeling approaches like that of Woodward (71) and Sloman (62), I can make two brief comments. First, my later discussion of counterfactual conditionals and backtracking is intended to vindicate talk of the ‘intervention counterfactuals’ that are invoked by causal modeling approaches. Thus, I see my account as complementing causal modeling approaches to causation. Second, my primary reason for preferring my account of the metaphysics of causation to any account based on the causal modeling approach is that I believe my account can help to elucidate why causation appears to be future-directed and why there is no genuine causal backtracking, whereas causal modeling approaches typically build these features of causation into their models, leaving them unexplained.

13.2 The Orthodox Metaphysics of Culpable Causation

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

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

13.2.1 Close Fit to Psychological Data

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

22 See Weslake (68) for an explanation of why this is so.
- (Irreflexivity) The rocket launch was not a cause of the rocket launch.
- (Asymmetry) The rocket launch was not a cause of the match being struck.
- (Preemption) The bystander’s decision to launch the rocket was not a cause of the rocket launch.

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

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

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

13.2.1.1 Irreflexivity

Lewis (44) advocates irreflexivity, stating that the cause “C and [effect] E must be distinct events—and distinct not only in the sense of nonidentity but also in the sense of nonoverlap and non-implication. It won’t do to say that my speaking this sentence causes my speaking this sentence; or that my speaking the whole of it causes my speaking the first half of it, or vice versa; or that my speaking of it causes my speaking it loudly, or vice versa.” (5), p. 78

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

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

\[^{23}\text{Notice that Lewis cleverly frames the issue in terms of what would be wrong to say, which permits the interpretation that it might be merely pragmatically misleading rather than explicitly false that events cause themselves.}\]
event and regardless of the laws of nature.

Not only is it acceptable to model self-causation as trivial causation rather than as a lack of causation, there is a good reason for doing so. Existing orthodox accounts already need explanatory pragmatics to account for why we do not cite causes that occur a trivially short amount of time before the effect, and these pragmatics automatically cover the case where the trivially short amount of time is zero. For illustration, suppose Jill is sleepy and goes to bed early in the evening and stays in bed until the late morning without anything remarkable happening. Let $E$ be Jill's being asleep in bed at exactly midnight. What are the causes of $E$? One of the causes is $C$, the fact that she is asleep in bed exactly $10^{-50}$ seconds before midnight. $C$ is not the kind of cause one would normally cite when providing a causal explanation of $E$ because on the time scales relevant to a causal explanation of human behavior it amounts to little more than a restatement of the event to be explained, but it does count as a cause according to the rules provided by prominent accounts so long as they permit reference to the brief events I described. Whatever story one employs to explain away the disutility of citing $C$ will likely extend to self-causation because in the limit as time $t$ approaches midnight, the event of Jill being in bed at $t$ becomes ever more useless for explaining $E$. If we take the orthodox approach and require that causation is irreflexive, then we in effect explain the wrongness of citing her condition at midnight as a cause of $E$ in two different ways depending on the fine difference between the state at precisely midnight and states arbitrarily close to midnight. At precisely midnight, her condition is not to be cited as a cause of $E$ because it is false that $E$ causes $E$, but at any moment just before midnight, her condition is not to be cited as a cause of $E$ purely on the pragmatic grounds that it is informationally unhelpful given typical human concerns. If we ignore the orthodox approach by adopting the unconventional hypothesis that events do cause themselves, we can say that pragmatics governs the wrongness of citing her condition throughout, so that there is not a discontinuity in the nature of the explanation that depends on the fine distinction between midnight and just prior to midnight.$^2$

13.2.1.2 Asymmetry

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

$^{24}$ Astute readers might want to counter that the asymmetry of causation requires that Jill's condition at an arbitrarily small time after midnight count as not a cause at all, so that we are left with a discontinuity regardless, but this can also be explained away, as discussed in the next subsection.
Another instructive illustration of orthodox analysis is its treatment of preemption. The decision of the bystander, $D$, to launch the rocket was a cause of the rocket launch, $E$, in the sense of being one of the events that played a part in the overall physical development of the world toward the launch. It also raised the probability of the launch. Suppose we accept, contrary to received wisdom, that $D$ is a genuine cause of $E$. We can explain why people have the intuition that $D$ is not a cause as follows: In the vast majority of cases, when there is causation from an event $C$ to a later $E$, there is a continuous physical evolution of the world from time $t_C$ to $t_E$ such that whatever difference $C$ eventually makes to $E$ is delivered by way of some physical differences in the intervening times. In our world, so far as we can tell, there are no nomic connections that leap over spans of time. Furthermore, because we often glean useful information about how the world works by observing patterns and tracing back from $E$ through whatever physical patterns we construe as causal processes, we place a lot of practical importance on those causes that can be found by tracing back in time from $E$. If all that is correct, we have a simple explanation for why we do not identify the bystander’s decision as a cause of the launch: The usual pattern of features we would expect if the bystander were causing the rocket to launch did not occur. We would expect things during that time span to exemplify a physical transition throughout the stages of a kind that is recognizably causal in the sense of matching what we think of as prototypical cases of a decision like $D$ leading to a rocket launch. Nevertheless, nothing about the lack of the right kind of process prevents us from claiming correctly that $D$ was a genuine cause of the launch although it is not recognized as such by folk judgment because it did not leave the usual indications that we use to identify causes. Of course, one could complain that identifying $D$ as a genuine cause does not capture the relevant notion of cause that the orthodox analyst is seeking, but this is exactly the tenet being questioned. Why is *that* notion of cause the one that needs to be enshrined in the metaphysics as genuine causation rather than some more liberal notion whose lack of psychological salience is explained away? The orthodoxy’s only answer is that folk do not cite $D$ as a cause when presented with such scenarios and that folk intuitions are the touchstones of adequate analysis. Because the orthodoxy permits some recalcitrant folk intuitions to be explained away, that calls into question how principled the egalitarian notion of cause really is.

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

One feature that makes the orthodox analysis of causation a project in metaphysics rather than armchair psychology is that a proper analysis is required to provide a principled account of what is common to all cases of causation. Suppose a psychologist offers a theory of causation consisting of a list of 8 exemplars of the cause-effect relation and 15 exemplars of the lack of a cause-effect relation. The theory says $C$ is a cause of $E$ if and only if the situation where $C$ and $E$ happens is closer to one of the positive exemplars than to any of the negative exemplars, closeness being judged by one’s own intuitive off-the-cuff assessment of similarity. A theory of this form might make for an interesting psychological theory and might even accrue empirical support if our causal reasoning is based less on rules than a pattern-matching capacity. But from the perspective of metaphysics, it fails to capture what is similar in all the cases of causation in an interesting way. Such theories come across as merely fitting the data, whereas the metaphysician is interested in a theory based on principles, something more closely resembling necessary and sufficient conditions.

13.2.3 STRICT STANDARDS FOR ACCOUNTS OF CULPABLE CAUSES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

13.4.1 Unclear Motivation

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

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

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

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

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

There is undoubtedly some benefit to having a unified theory of causation simultaneously honed to serve some metaphysical role as well as to account for why we have our shared body of intuitions about culpable causes. Such a theory would provide some valuable conceptual economy. But the relevant question is whether the gain in conceptual economy is worth the loss in conceptual optimization. A screwdriver made out of a carrot would have clear benefits; it would be lighter than an ordinary screwdriver and you could eat it if hungry. But given the obvious tradeoffs, it is hard to believe an engineer could design a carrot-screwdriver that would not be significantly outperformed by just having a metal screwdriver and an ordinary carrot separately. Conceptual economy is worth something, but not much. If a theory were to invoke a sizable number of different versions of our causal concepts without a clear enough account of how they are related, that would be a good reason for complaint. Having two or three kinds of causation and a story about how they fit together hardly strains our cognition. But the cost of replacing two concepts optimized toward different ends with a single causal concept that is optimized toward both simultaneously is significant. Barring a stroke of fortune, the complexity of our psychology of causation demands tradeoffs between the degree of fit with common sense intuitions and the simplicity of the rules governing the application of causal concepts. The new-fangled orthodox analyst already admits this, in that the whole point of the egalitarian notion of cause is to idealize away opinions that result from explanatory pragmatics for the sake of a simpler account of causation. What’s more, the difficulty in achieving a good fit with folk judgments while being simple and comprehensive is easily measured by the vast volume of material written on the subject of culpable causation. Once the practical necessity of these tradeoffs is accepted, there is room for different accounts to trade off the fit in different ways for different purposes. What is peculiar about the orthodox analysts’ take on causation is their frequent insistence that there is one right
way to optimize the concept.\footnote{Hitchcock (\textcopyright1) discusses numerous examples of such pseudo-debates.} I suspect that what explains this curiosity is that orthodox practitioners conceive of conceptual analysis more as an activity of conceptual exploration and discovery rather than conceptual engineering and construction.

### 13.4.2 Causation Is a Scientific Concept

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

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

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

The challenge for the orthodoxy is to explain what makes causation special in a way that requires that its \textit{strict} conceptual analysis must be moored closely to folk opinions about causation while the conceptual analyses of all the various species of causation need only match folk opinions in the loose way that is uncontroversially acceptable in science. It does no good to cite the greater metaphysical significance of causation, for empirical analyses of causation are required to deliver a principled, \textit{strict} analysis of causation as well. All an empirical analysis lacks is that the aspects of our folk conception of causation irrelevant to the explanation of effective strate-
gies are delegated to the psychology of causation where they are given a relaxed treatment. If the empirical analysis of causation and the empirical analysis of the psychology of causation succeed together at providing a complete scientific explanation of effective strategies and a complete scientific explanation of why we have the naive causal concepts we have, on what grounds will the orthodox defender argue that these explanations do not tell us everything we need to know about causation?

13.5 Changing the Topic

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

13.6 Summary

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

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

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

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

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

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

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

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

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

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

Because one of the main reasons we have a notion of culpable cause is that it aids our discovery of promotion or prob-influence relations—a hypothesis I suggested in §8.2—we should expect this function to reveal itself in our judgments. It will turn out in §4.4 that there are discrepancies between what we would judge culpable if we cared only about whether that one particular instance of $C$ affected the probability of that one particular instance of $E$ and what we would judge culpable if we cared more about the discovery of prob-influence relations that apply to more general circumstances. When such discrepancies appear, according to my theory, we should expect our instinctive judgments concerning culpable causes to track the latter because such thinking would have greater practical utility.

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

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

An event is a culpable cause of $E$ iff it successfully induces $E$.

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

This guiding principle is one variant of the hypothesis that singular causation can be adequately understood in terms of probability-raising processes. This should not be surprising because such theories are motivated primarily by the goal of incorporating (1) some sort of production or process or mechanism with (2) some sort of counterfactual dependence or difference-making or probability-raising. Because the metaphysics of causation I have presented represents...
To begin the investigation of this guiding principle, I will first impose a simplifying assumption, second comment on ‘induces’, and third comment on ‘successfully’.

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

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

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

I invite readers to interpret ‘induces’ liberally enough to include models of causal tendencies expressed in terms of forces or hastening or intentions. For example, there is a sizable literature in psychology based on the suggestion of Talmy (55) that many of our intuitions about causation can be effectively modeled in terms of our conception of force vectors. Wolff and Zettergren (69) report that a force-based approach successfully predicts a range of causal judgments regarding material objects. For example, if a motorboat is attempting to go away from a buoy but a strong wind blows it back until it hits the buoy, people will say the wind caused the boat to strike the buoy. Also see Wolff (70). The pronouncements of this “force dynamic” model of causation, I believe, overlap enough with the pronouncements one gets from a well-designed model based on difference-making in order to justify the following claim. If it is useful for a creature to possess the psychological faculties described by one of these two models—the “force dynamic” model or the difference-making model—it is useful for a creature to possess the psychological faculties described by the other. Similar comments apply to cases where someone hastens the occurrence of an effect that would have happened later without the action and to cases where someone acts intending for a certain effect to occur. The scenarios where these models disagree are important for debates in psychology, but I will not be concerned with their differences because the toy theory is only intended to establish a fairly reliable link between promotion and our assessments of culpability, not to insist that people’s reasoning about causal tendencies must closely match probabilistic relations.
each candidate, \(c_i\), under some not-too-convoluted coarse-grained description, \(C_i\). For brevity, I will use the expression “\(c (\text{as } C)\)” to refer to the fine-grained event \(c\) under the coarse-grained description \(C\). In order for it to be connected to the metaphysics, though, the event also needs to be thought of as a contrastive event, \(\tilde{C}\), which comports with the observations of §4.8 that we often tend to use implicit contrasts when thinking of culpable causes. In all cases that we need to consider, the contrastive event is intended to be a contrastivization of the coarse-grained event having its background conditions filled in with a reasonable contextualization of \(C\)’s actual environment at the same time as \(C\). For brevity, I will use the shorthand “\(c (\text{as } C \text{ qua } \tilde{C})\)” to signify that \(c\) has been coarse-grained as \(C\) and contrastivized as \(\tilde{C}\).

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

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

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

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

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

14.2.1 SALIENCE

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

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

Some evidence exists that moral categories play a role in our selection of which events potentially count as causes, for example, Alicke Alicke (4), Knobe and Fraser (56), and Driver (12) (13). This would be surprising in a model of moral judgments where step one is to ascertain which events count as causally relevant without any appeal to

\[\text{See Maudlin (47) and the discussion of default and deviant states in Hall (23) and Hitchcock (13).}\]
morality, and step two is to apply moral principles to assess those events for moral culpability. Although investigation of the role of morality in people’s identification of culpable causes is in its infancy, the claim that our beliefs about morality play a role in whether some chosen event counts as a cause would not be surprising given my theory of causation. Because the concept of culpable cause is parasitic on the notion of promotion, culpable causes inherit the contrastivity of promotion. And, as noted in §4.8, the default contrasts people use in assessing causal promotion include what people believe is normal or what they believe should happen. We can think of “what should happen” as what typically happens, or as what will happen if things work as they are intended or designed to function, or as what the law or morality dictates. All these senses of ‘what should happen’ can play a role in identifying candidate causes.

For example, when determining why a particular bridge collapsed, we tend to sift through events that differ from the norm in one of these senses. We might flag the existence of an unusually heavy load as a candidate cause just because it is atypical. Or we might flag the failure of a certain joint to maintain rigidity as a candidate cause because the purpose for which it was installed was to hold its beams rigidly together. Or we might flag the inspector’s negligence because he was legally obligated to check the joints and morally obligated to make a good-faith effort. Actions people take in accordance with the law and morality are ceteris paribus less likely to be salient because routinely considering them would usually result in an unmanageably large number of candidate causes.

Another factor governing whether an event counts as salient is how broadly it is coarse-grained. The coarse-graining is often selected by some sort of default conception of an event, but we also have the ability to select a coarse-graining as salient in a more sophisticated manner. Imagine observing a person who is the subject of a psychological experiment. The subject attends to an unlit button on a panel; the button lights up with a green color; and the subject responds by pressing the button. It is natural to conceive of the situation as one where the lighting of the button caused the person to press it or where the lighting of the button as green caused the person to press it. One would not normally think of the cause as “the button lighting as either green or yellow” because there is no reason to suppose the button can light up as yellow or that a yellow light would induce the subject to press the button. However, if you are told the subject was instructed to press the button when and only when the light appeared as either green or yellow, and you see the button turn green and then the person pressing it, it would be reasonable for you to describe the cause as “the button lighting up as green or yellow.” That description is appropriate because you know the most informative description of what is promoting the person to press the button is its lighting up as either green or yellow. It is reasonable to select this “green or yellow” contrastivization to inform one’s selection of a salient candidate cause even though nothing in this particular case prevents one from accurately describing the cause more narrowly as “the button turning green.” (Communication of the intended contrast also plays a role here.) This feature accords with my contention that our intuitions about culpability are often tuned in order to be useful for conveying information about promotion. Unlike Yablo’s (72), (73) principle that “causes must be proportional to their effects,” however, culpable causes in my account need not be coarse-grained in a maximally informative way.
There is also quite a bit of flexibility and lack of specificity not only in how we select some contrasts as the appropriate ones for defining the candidate causes but also in how broadly to coarse-grain the background conditions. The key idea motivating culpability, is that we have some conception of what it is for an event to be generally good at bringing about the general kind of effect that $E$ represents. To cash out this idea appropriately, the background conditions implicit in the contrastive events need to be both broad enough and specific enough to capture an ordinary understanding of the general conditions under which events occur. For example, in identifying the culpable causes for a campfire, we typically seek promoters of campfires by considering contrastive events that range over a wide range of earthly environments. But the relevant extent of the contrastive events does not extend to include conditions present in deep space or at the bottom of the sea. The appropriate degree of generality is something that the toy theory leaves as a rather flexible parameter.

14.2.2 IRREFLEXIVITY

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

14.2.3 ASYMMETRY

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

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

In §7.45, I noted the existence of non-local partial influence. Although I believe that provides a legitimate, albeit esoteric, sense in which two simultaneous events can non-trivially promote each other, I believe it is far enough removed from the way people ordinarily think of causation, to disregard it when theorizing about the psychology of causation. People might occasionally employ reasoning that corresponds to pseudo-backtracking connections, but because non-local partial influence is only exploitable in the way described in §9.7, it makes sense for people, upon a modest amount of reflection, to interpret non-local partial influence as not being genuinely causal, even though according to my theory it really is.

### 14.2.4 significant promotion

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

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

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

### 14.2.5 CAUSAL GROUPING PRINCIPLES

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

In chapter 2, the concept of overdetermination was discussed with regard to multiple fine-grained events determining the same event. However, there is an altogether different notion of overdetermination that pertains to culpable causation. Regarding
culpability, **overdetermination** occurs when multiple distinct events are culpable as a group for some effect, and also individually. I will define overdetermination only for the simplest example where there are two salient events.

Two distinct existing events \( c_1 \) (as \( C_1 \)) and \( c_2 \) (as \( C_2 \)) are overdetermining culpable, causes of \( e \) (as \( E \)) if and only if all of the following hold:

1. \( p_{c_1 \& c_2}(E) \gg p_{\neg c_1 \& \neg c_2}(E) \)
2. \( \neg \left( p_{c_1 \& c_2}(E) \gg p_{\neg c_1 \& \neg c_2}(E) \right) \)
3. \( \neg \left( p_{c_1 \& c_2}(E) \gg p_{\neg c_1 \& \neg c_2}(E) \right) \)
4. \( p_{c_1 \& c_2}(E) \gg p_{\neg c_1 \& \neg c_2}(E) \)
5. \( p_{\neg c_1 \& \neg c_2}(E) \gg p_{c_1 \& c_2}(E) \)

where ‘\( \gg \)’ means ‘significantly greater than’, as discussed in §14.2.4, and the obvious contextualizations are employed.

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

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

Two distinct existing events \( c_1 \) as \( C_1 \) and \( c_2 \) as \( C_2 \) are joint culpable causes of \( e \) as \( E \) if and only if all of the following hold:

1. \( P_{c_1 \& C_2}(E) \gg p_{\neg c_1 \& \neg C_2}(E) \)
2. \( P_{c_1 \& C_2}(E) \gg p_{\neg c_1 \& C_2}(E) \)
3. \( P_{c_1 \& C_2}(E) \gg p_{c_1 \& \neg C_2}(E) \)
4. \( \neg (p_{c_1 \& \neg C_2}(E) \gg p_{\neg c_1 \& \neg C_2}(E)) \)
5. \( \neg (p_{\neg c_1 \& C_2}(E) \gg p_{\neg c_1 \& \neg C_2}(E)) \)

where the obvious contextualizations are employed.

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

My reason for mentioning these grouping principles is that people are capable of judging culpability based on different ways of grouping events, and concepts like overdetermination and joint causation provide a richer picture of the underlying promotion relations, especially when the promotion relations occur in complex combinations. For example, one could have a situation where a group of five events is together culpable, for an effect \( E \). To evaluate culpability, for each individual event, one would consider the default background conditions where the occurrence of the other four events is held fixed. But one could also consider other non-standard background con-
Each 0.7kg mass (in the presence of the other) promotes the balance being tipped right. Neither mass (in the absence of the other) promotes the balance being tipped right.

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

### 14.3 Shortcomings of Culpability

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

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

14.3.2 OVERLAPPING CAUSATION

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

14.3.3 PROBABILITY-LOWERING CAUSES

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

---

31 See also, Schaffer (55).
The Psychology of Culpable Causation

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

14.3.4 TRUMPING

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

$$F_X = k_w \frac{w_Xw_Y}{(r_{XY})^2} + k_s \frac{s_Xs_Z}{(r_{XZ})^2},$$

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

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

$$F_X = k_s \frac{s_Xs_Z}{(r_{XZ})^2} = k_s \frac{(+1)(+1)}{(2\text{ nm})^2}$$

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

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

14.4 Culpability$_2$

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

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

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

---

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

33 Recall from §7.3.1 that the prominent foreground of $G$ is the region of the arena where $G$ and $\sim G$ differ significantly, and the prominent background is the complementary region where they do not differ significantly.
For an illustration, consider the case of overlapping causation from §14.3.2, where
the left machine is 99% accurate and can hit to the left of the canvas, and the right
machine is 1% accurate and can hit to the right of the canvas, and the machines do
not significantly interact with each other. The chosen actual effect \( e \) is a splattering of
paint on the canvas and to the right of the canvas but not to the left. We stipulate our
coarse-grained effect of interest, \( E \), to be the existence of some paint on the canvas.
Construed as a contrastive event, the firing of the right machine is
\( \sim C_r \) (\( C_r \); \( \sim C_r \)),
which fixes \( G \equiv (\sim G; \sim G) \) for region R. In this example, \( G \) and \( \sim G \) are very nearly alike
except that \( G \) is far more likely than \( \sim G \) to instantiate a paint splatter on the right side
of the wall. Because \( C_r \) and \( \sim C_r \) very nearly agree on the likely motion of the left paint
ball, everything that happens with the left paint ball is excluded from what is (trivially)
promoted by (the prominent foreground of) \( G \). What \( \sim C_r \) is “trying” to promote is not
\( E \), paint on the canvas, but paint being somewhere on the canvas and/or on the wall
to its right rather than no paint in that region.

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

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

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

The final step is to pair these two contextualized events together to form
\( \sim E \equiv (E_1, E_2) \),
which represents everything \( \sim C_r \) successfully promoted for region R. To evaluate
whether \( C_r \) successfully induced some chosen \( E \), we now only need to consider
the degree to which \( \tilde{E} \) (trivially) prob-influences \( E \). We can always read off of \( \tilde{E} \) the degree
of prob-influence for any plain coarse-grained event \( E \) in \( R \) as follows. The degree to
which \( \tilde{E} \) prob-influences \( E \) is equal to the proportion of \( \overline{E}_1 \)'s members that instan-
tiate a member of \( E \) (in the correct region) minus the proportion of \( \overline{E}_2 \)'s members
that instantiate a member of \( E \) (in the correct region). The degree to which \( \tilde{E} \) prob-
influences \( E \) is by construction equal to the degree to which \( \tilde{C}_r \) successfully promoted
\( E \) and thus is equal to the degree to which \( C_r \) successfully induced \( E \). If this degree of
successful inducement is significantly positive, then \( C_r \) counts as a successful inducer
of \( E \).

In order to fill in the gap left in this procedure—removing appropriate members from
\( \overline{E}_2 \)—there are at least two guiding principles we should apply. The first principle in-
volves removing aspects of \( \overline{E}_2 \) that are not of the right kind to be prob-influenced by
\( \tilde{C}_r \). The second principle involves removing aspects of \( \overline{E}_2 \) that can be attributed to
other independent causes. We remove an aspect of \( \overline{E}_2 \) by stripping out members of
\( \overline{E}_2 \) to equalize the probabilities that \( \overline{E}_1 \) and \( \overline{E}_2 \) fix for that aspect. I will illustrate both
principles with examples.

The first way to tell whether some aspect of \( \overline{E}_2 \) should be removed is to examine what
kind of effects \( \tilde{C}_r \) promotes for region \( R \). Suppose for example that a particular
location on the canvas and wall is lit by several spotlights that flicker on and off every
now and then with \( \tilde{C}_r \) not prob-influencing anything about the lights. Because \( \tilde{G} \) does
not prob-influence anything regarding amount of light striking the wall or canvas, we
should adjust \( E_2 \) (to match \( \overline{E}_1 \) with regard to its pattern of luminosity) so that \( \tilde{E} \) does
not prob-influence the amount of light striking the wall. In this way, we render \( C_r \),
not culpable for the canvas being lit the amount that it is, for the wall being at room
temperature, for the existence of a roach at a particular location on the floor, and so
on.

A special case of this principle involves transferring the prominent background of
\( \tilde{G} \) to \( \tilde{E} \), as can be seen in our current example of overlapping causation where \( \overline{E}_1 \)
fixes a very high probability for the particular pattern of paint on the wall to the left
side of the canvas (which came from the left machine). Unamended, \( \overline{E}_2 \) would fix a
very low probability for any particular splotch of paint because \( \neg \overline{C}_r \) leaves open the
full range of possibilities for where the left machine's paint can land. Unamended, \( \tilde{E} \)
would thus count as successfully promoting the splotch of paint on the left, which
disagrees with our judgment that the firing of the right machine was not one of the
causes of the left machine missing the canvas. To resolve this problem, it is reason-
able to strip out members of \( \overline{E}_2 \) to make the prominent background of \( \tilde{E} \) match the
prominent background of \( \tilde{G} \). The prominent foreground of \( \tilde{G} \) is the subregion of \( R \)
that includes everywhere the right paint ball could have landed, and its prominent
background is everywhere else, including the actual location of the paint splotch on
the left. The technical implementation of the solution is to discard any members of
\( \overline{E}_2 \) that disagree with \( G_E \) in the prominent background of \( \tilde{G} \), and then renormalize

\[34\] The relevant kinds here exclude any kinds that are too difficult for people to cognize.
its probability distribution. By doing so, the prominent background of \( \bar{E} \) will be the same region as the prominent background of \( \bar{G} \).

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

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

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

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

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

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

---

35 One could at this stage incorporate additional considerations related to causal grouping—overdetermination and joint causation—but unfortunately I have had to abbreviate this presentation of the toy theory.
section is that it does not work nearly as well when the contrast in the candidate cause is likely to interact with stuff in the environment and leave traces in the fine-grained effect. Unfortunately, I will have to forego how to refine the method further.

It is a good exercise at this stage to consider other cited shortcomings of culpability₁ in order to see how culpability₂ helps secure better agreement with our pre-theoretical judgments of culpable causation. In the example from §4.3.4, the right machine fires periwinkle paint balls and the left fires some other color. The firing of the left machine will not count as a successful inducer of \( E \). Although \( C_r \) promoted paint on the canvas, it did not promote the existence of periwinkle paint on the canvas, and yet periwinkle paint is the only color of paint on the canvas and thus was the only color of paint that was represented in the constructed \( \sim E \).

In the example from §4.3.6, the firing of the right machine lowers the probability of paint on the canvas, yet we still think of it as culpable for the existence of paint on the canvas. We can now make sense of this judgment. Even though \( C_r \) makes paint on the canvas much less likely, it significantly promoted the probability of the precise pattern that happened to appear. \( C_r \) was successful at inducing the more finely grained effect and so was culpable for \( E \).

Given my previous suggestion that the reason we have a notion of culpability is that it allows us to more quickly infer promotion relations, it is worth considering why it would benefit us to have intuitions that match culpability₂ rather than culpability₁. I certainly do not think our psychological mechanisms for attributing singular causation implement the precise form of my exegesis of culpability₂. However, it is a plausible hypothesis that we reckon culpable causes by scouring the evidence included in the environment of the fine-grained effect and piecing together which candidate causes were responsible for which aspects by attributing each chosen aspect to a candidate cause (or candidate group of causes) when it is the only candidate that could have promoted that aspect.

In many circumstances, culpability₂ is not too much harder to assess, and it is often much more responsive to the observed evidence than culpability₁. As discussed in §8.2, recognizing which of the two machines is culpable for the effect often allows one to make good estimates about how much each machine individually promotes the effect. In circumstances where the paint-ball-firing machines barely influence one another, one can easily gather statistics on how often the splotch of paint spreads significantly toward the right and thereby infer the fraction of times that a \( C_r \) event was culpable for the existence of paint on the canvas, \( E \). That, in turn, provides a good estimate for how likely the machine on the right would place paint on the canvas when operated alone.

### 14.5 Shortcomings of Culpability₂

The primary shortcoming of culpability₂ is that we sometimes rule out a candidate cause because it does not successfully deliver its inducement to the effect through
an appropriate process. In technical terms, culpability\_2 does not properly account for ‘fizzling’, a term from Schaffer\textsuperscript{(55)}.

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

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

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

14.5.1 SAVED FIZZLES

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

The next two subsections discuss cases of preëmption, that are also special cases of saved fizzes.
14.5.2 EARLY CUTTING PREÉMPTION

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

14.5.3 LATE CUTTING PREÉMPTION

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

14.6 Culpability

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

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

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

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

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

14.7 Culpability\(_4\)

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

An actual event \( c \) (as \( C \) qua \( \tilde{C} \)) is culpable\(_4\) for an actual event \( e \) (as \( E \)) iff there is a chain of culpability\(_3\) relations running from \( c \) to \( e \).
Because there are no hard and fast rules about which intermediate events count as salient inducers or how they are to be rendered as contrastive events, culpability is sensitive to our choices of how to abstract away from the fundamental material layout. By being extremely permissive about event salience, one can achieve an extremely long chain of very finely grained events that are only slightly apart in time. Being extremely permissive as a general policy would result in so many culpability relations that culpability would have little utility. So, our employment of culpability needs to be restricted to a suitably limited class of salient events if we want it to do interesting psychological or explanatory work.

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

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

Let (as ) be the event of the train crossing the road, and let (as ) be the activation of the switch for the left track rather than the right track. is culpable for because significantly promotes the train moving along the left track, which in turn significantly promotes the train crossing the road, . Note that the intermediate event does not use the contrast that is fixed by but is chosen by reckoning salient contrasts at the intermediate time. The switching event is arguably not culpable for because does not promote the slightly coarse-grained version of . After all, the probability of the train’s reaching the road is the same whether the switch is thrown or not. As with most judgments of culpability, it is possible in principle to argue that there is some just barely coarse-grained version, , of the fine-grained effect, , such that the switching, qua , promotes and is thus culpable for . Such an , though, would need to include the kind of fine details about the likely character of the train had it taken the left track versus the right. For example, there might be more flies near the left track, so that a train taking that route would tend to displace more flies. If such a finely grained construal of the effect were to be countenanced as part of our standards for judging culpability, there would be many more culpable causes than we actually judge. Thus, we can set aside (as too deviant) such a finely grained construal of the
People who are fully aware of what happened in this scenario will be likely to say that the switching event was not one of the causes of \( E \), largely because it is clear that the switching makes no significant difference to how \( E \) comes about. Their judgments match what is culpable\(_3\). However, if the example is altered slightly, people’s judgments will likely match what is culpable\(_4\):

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

Our intuitive judgment in such cases is that the switching event was one of the causes of the train successfully crossing the road. Again, the switching is culpable\(_3\) for the train’s traveling along the left track, which is later culpable\(_4\) for the train’s making it to the road crossing, \( E \); thus, it is culpable\(_4\) for \( E \). And again, the switching would not be culpable\(_3\) for \( E \) because at the time the switch is thrown, the chance of the train’s eventually reaching the road crossing is the same whether it goes along the left track or the right track. Because the switching event does not prob-influence anything concerning the tree, the assessment of what \( \tilde{C} \) successfully promoted is not supposed to change (according to the simplistic method for constructing contrastive effects discussed in §1.4.4). This pair of examples shows that sometimes our common-sense judgments of culpability match culpability\(_3\) but not culpability\(_4\), and that sometimes they match culpability\(_4\) but not culpability\(_3\).

When people ask, “What are the causes of \( E \)” they usually do not distinguish between these two different kinds of culpability. But once it is apparent that the toy theory posits these two distinct versions of culpability as guides to our (often presumed to be univocal) implicit notion of culpable cause, it follows that the toy theory has a conflict in the technical sense introduced in §1.8. The toy theory tells us that one good rule of thumb for assessing culpability is that an event is culpable for \( E \) iff it is culpable\(_1\) for...\

---

\(^{36}\) Because this chapter is only sketching a toy theory of the psychology of culpable causation, not every deficiency can be discussed, but I believe interested readers would benefit from exploring how the motivation I have suggested for distinguishing culpability\(_4\) from culpability\(_3\) by formulating a more sophisticated scheme for constructing the contrastive effects than the one that I assumed when extending the considerations in §1.4.4 to handle the temporally extended process leading from the cause to the effect. Specifically, the scheme I presented does not take into account that one’s chosen contrast in the cause ought to interact with stuff in the background to help generate the proper contrast to use for representing the effect. If one does so, it may be possible to render the switching event culpable\(_1\) for the train’s making it to the road crossing, though I suspect people do not reason very clearly about sequences of merely hypothetical interactions beyond simple cases.
It also tells us that another good rule of thumb for assessing culpability is that an event is culpable for \( E \) iff it is culpable\(_4\) for \( E \). Because there are realistic circumstances where an event is culpable\(_4\) without being culpable\(_3\), the theory provides a conflicting account of which events are culpable. Furthermore, nothing in the toy theory ameliorates this conflict by specifying conditions that adjudicate which version of culpability should supersed e the other. According to empirical analysis, these genuine conflicts do not imply that the toy theory is incoherent, nor do they imply that one of the two rules of thumb needs to be rejected as fatally flawed. On the contrary, both versions of culpability have limitations as guides to our cognition of culpable causation, and each offers different benefits. When investigating the psychology of culpable causation at a fairly high level of abstraction, as the toy theory does, it is acceptable to employ relaxed standards where these kinds of conflicts do not need to be ameliorated with an explicit rule. Insofar as we are just sketching the outlines of a full psychological account, we do not need to specify in every possible instance whether culpability\(_3\) or culpability\(_4\) is the “correct” account of culpability. And insofar as we are investigating the metaphysics of causation, we do not need an account of culpability at all. The conflict in the toy theory that exists by virtue of its not privileging culpability\(_3\) over culpability\(_4\) or vice versa does not count as a reason to reject the toy theory qua toy theory.

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

I have already discussed why it is reasonable for people to have intuitions about culpability that match culpability\(_3\). Now I would like to cite a few reasons why it is reasonable for people to have intuitions about culpability that match culpability\(_4\). For one, fixing plausibly obeys unidirectional transitivity and continuity, as discussed in §4.9 and §4.10, and to a great extent, relations of culpability serve as cognitive proxies for promotion relations. So, it is often convenient for us to think of culpability relations as being continuous and transitive just like the promotion relations they approximate. As Example 14 and the two examples from §4.9 demonstrate, it is not correct to think of our intuitive conception of culpability as transitive, but in a wide range of situations, it is convenient to think of \( C \) as successfully inducing \( E \) by virtue of successfully inducing an intermediate event, which successfully induces another intermediate event, and so on until \( E \) occurs. Because the metaphysics of promotion among contrastive events is too complicated for people to manage cognitively, it is understandable that people approximate the unidirectional transitivity of promotion by largely ignoring the background conditions and just thinking of causation as occurring by virtue of a localized chain of events or a localized continuous process.

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

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

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

---

37 For similar observations, see Edgington (16), Kvart (19), Northcott (52), and my discussion of “infection by culpable causation” in the supplementary material I have provided concerning Morgenbesser’s coin.
the lake, then it is reasonable to suspect that a chemical from the factory is culpable for the condition of the water. Because we have previously learned that waterborne chemicals are sometimes culpable for reduced bird reproduction and because we know just by looking that the factory is plausibly culpable for some sort of effect on the watershed, we are justified in inferring that there is a reasonable possibility that the factory is culpable for the failure of the ducks to reproduce. This can justify restricting the testing to chemicals used in the factory instead of testing for the full array of epistemically possible chemicals in the lake. If it were illegitimate to identify potential causes by using what we know about chains of culpability, we might waste time testing other possible sources. For example, if we can be sure that the chemicals stored in some nearby warehouse never left the warehouse, we can be reasonably sure that they are not culpable for any effects on the water supply, and thus reasonably sure that they are not culpable for the canvasbacks’ troubles. This indicates that it is likely unnecessary to test the water for these chemicals.

One of the consequences of having both culpability and culpability is that many questions about culpability that might initially seem straightforward become extremely messy. An exemplary complicated case is Hesslow’s thrombosis example. Taking a birth control pill regularly is a promoter of thrombosis by virtue of its direct role as chemical in the body. But the birth control pill is also an inhibitor of pregnancy, which itself is a promoter of thrombosis. So, there are two routes by which thrombosis is probabilistically influenced. For the sake of discussion, we can modify the example to have them approximately cancel each other out over the course of time, so that taking the pills on the whole has no net probabilistic influence on thrombosis. Imagine that some woman takes the birth control pill, does not become pregnant, and is not afflicted by thrombosis. Is her consumption of the pills one of the causes of her being free of thrombosis? It might seem that the pills cannot be culpable for her failure to contract thrombosis because they do not prob-influence thrombosis. It might also seem that the pills are culpable for her not getting thrombosis because there are many chains of successful promotion that run from her taking a given pill to her lack of thrombosis at later times. So, the two notions conflict in their attribution of causal culpability. A univocal assessment is made even more difficult when we take into account that taking pills for a full year consists of many localized events: the daily occurrences where she ingests a single pill. It is plausible that many of these events exert different degrees of promotion and inhibition through different intermediate mundane events. Furthermore, whether an event is culpable depends on which events are permissible for employment in chains of culpable causation. Remember that if we identify salient events liberally, allowing all sorts of non-standard contrasts and coarse-grainings, just about any event will count as culpable for her not having thrombosis. So, the relevant culpability would have to be restricted to some appropriately salient events in order to match our psychological judgment that there do not exist a vast multitude of thrombosis preventers. But we do not have clear intuitions about how to break down vast causal networks (like those present in the daily operation of the human body) into relevant component events. In summary, it is safe to say that in complicated interactions like those exhibited by the thrombosis example, it is difficult to make unequivocal statements about culpable causation that are well grounded in our practices of attributing culpability.
At this point, my presentation of the toy theory is complete. There are several more features of the toy theory that constitute additional evidence that the toy theory meshes well with my account of the metaphysics. The degree of support provided by each of the following sections to the metaphysical theory of causation is individually small, but I think each discussed feature helps to reinforce my contention that the toy theory of causal culpability is not concocted ad hoc to accommodate psychological data but is reasonably well motivated by the hypothesis that we have a conception of culpable cause because it provides a useful shortcut for learning about promotion. The following sections are meant as stand-alone commentaries and are not presented in any special order.

14.8 Uncaused Events

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

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

---

38 Remember that no events are altogether uncaused. Every fine-grained event trivially terminates itself, and every contrastive event trivially promotes itself.
ECHO. However, it makes sense to ignore states like \( s \) when judging fizzling because these kind of states are reliably produced by ALIGN and reliably result in ECHO. The unusual character of \( s \) is in practice inaccessible by observation of \( s \) alone, but because we have so many successful cases of promotion that run from ALIGN to FLIP to ECHO, it is clear that the development from FLIP to ECHO is no mere accident. Thus, we ought to incorporate into our rules for fizzling an exception clause to handle special cases like \( s \), where the intermediate structures needed for promotion are present but are hidden in the microstate.

### 14.9 Prevention

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

The factors that affect salience for prevention are much like those for ordinary cases of culpable causation. For example, consider some scenarios tested by Walsh and Slozman (57):

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

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

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

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

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

14.10 Double Prevention

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

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

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

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

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

### 14.11 Culpable Causation by Omission

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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


References


Index

actuality-focusing, 13
antecedent event, 12
backtracking counterfactuals, 55
bizarreworld, 3
causal asymmetry, 54, 55
chain, 51, 79
explanation, 84
irreflexivity, 54
mechanism, 59
process, 53, 77
causal grouping, 60, 99
causation culpable, 59
non-metaphysical aspects of, 91
orthodox metaphysics of, 53
singular, 12, 59
causation by omission, 88
cause culpable, 12
chain, causal, 51, 79
conceptual analysis
orthodox, 58
conditional nomic, 1
conflict, 81
conflict amelioration
delegation of, 58
consequent event, 12
continuity, 82
counterfactual, 41
mundane, 12
counterfactual dependence, 59
counterlegal, 59
culpability, 53
as proxy for promotion, 57
as tension in concept of, 59
culpability, 1
culpability, 2
culpability, 3
culpability, 4
culpable causation, 59
culpable cause, 42
difference-making, 59
double prevention, 87
empirical analysis of the non-metaphysical aspects of causation, 21
enabling cause, 83
event mundane, 13, 43
example
balance scale, 23
bridge collapse, 53
duck reproduction, 83
fired gun as double prevention, 85
fizzled rocket fuse, 77
lit panel button, 83
lottery ticket, 86
match striking, 72
neutrino-triggered bomb, 83
paint balls, 22
surfing, 85
thrombosis, 84
tiger, 89
track switching, 83, 85, 87, 89
positive cause, 88
preemption, 48
early cutting, 78
late cutting, 78
probability-raising, 59
production, 58
relaxed standards of theoretical adequacy, 84
salient promoter, 52
saved fizzle, 77
fizzle, 77
saved, 77
fizzling, 85
focal set, 62
force dynamic model of causation, 54
general causation, 42
managed inconsistency, 58
mechanism, causal, 55
minimal account of counterfactuals, 10
Minkowski space-time, 27
morality and causation, 52
Morgenbesser’s coin, 3
mundane counterfactual, 12
mundane event, 13
new-fangled orthodox analysis, 57
nomical conditional, 1
non-spatiality, 27
old-fashioned orthodox analysis, 35
omission, 88
orthodox (conceptual) analysis, 35
new-fangled, 57
old-fashioned, 35
overdetermination, 56
preemption, 48
preemption, 48
preemption, 48
primary event, 58
relaxed standards of theoretical adequacy, 84
track switching, 83, 85, 87, 89
tree falling on track, 83
exemplars, 55
explicitly true, 84
fizzle, 77
saved, 77
fizzling, 85
focal set, 62
force dynamic model of causation, 54
general causation, 42
managed inconsistency, 58
mechanism, causal, 55
minimal account of counterfactuals, 10
Minkowski space-time, 27
morality and causation, 52
Morgenbesser’s coin, 3
mundane counterfactual, 12
mundane event, 13
new-fangled orthodox analysis, 57
nomical conditional, 1
non-spatiality, 27
old-fashioned orthodox analysis, 35
omission, 88
orthodox (conceptual) analysis, 35
new-fangled, 57
old-fashioned, 35
overdetermination, 56
preemption, 48
probability-raising, 59
positive cause, 88
production, 58
relaxed standards of theoretical adequacy, 84
salient promoter, 52
saved fizzle, 77
<table>
<thead>
<tr>
<th>Term</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>significant promotion</td>
<td>65-66</td>
</tr>
<tr>
<td>simultaneous causation</td>
<td>65</td>
</tr>
<tr>
<td>singular causation</td>
<td>42-59</td>
</tr>
<tr>
<td>SM-conditional</td>
<td>24</td>
</tr>
<tr>
<td>space-time Minkowski</td>
<td>27</td>
</tr>
<tr>
<td>sufficiently specified antecedent</td>
<td>13</td>
</tr>
<tr>
<td>thrombosis example</td>
<td>84</td>
</tr>
<tr>
<td>top conceptual layer of causation</td>
<td>57</td>
</tr>
<tr>
<td>transitivity</td>
<td>82</td>
</tr>
<tr>
<td>trumping</td>
<td>71</td>
</tr>
<tr>
<td>underspecified antecedent</td>
<td>12</td>
</tr>
</tbody>
</table>