3

What Is a Cause?

Causes and Effects Are Events

In everyday language, causes and effects assume various roles. We say drugs cause addiction, sparks cause fire, and guns cause death (or at least bullets do). In each case, an object (drugs, sparks, guns) is the cause of a state (addiction, death) or an event (fire). Physical and emotional states can cause other states, as fear causes loathing, hunger causes suffering, or satisfaction causes tranquility. Events can cause other events, as one war causes another, or a strong wind causes a tree to fall.

It's harder to think of cases when we say a physical object causes another physical object. Parents cause their children, in a sense, and factories cause their products, but these statements don't sound quite right. The reason such cases are hard to think of is that we conceive of objects in a static way, as if they are fixed over an extended period of time, yet causal relations are enacted over time.

A causal relation suggests a mechanism unfolding over time that uses the cause (and possibly other things) to produce the effect (and possibly other things). So the notion of cause involves change over time, whether the time interval is short (as a light source causes a shadow), long (as the big bang causes the universe to expand), or intermediate (as the earth's rotation causes the day to turn into night). One general temporal constraint on causation is that effects cannot precede their causes.

Although cause suggests mechanism, it's a mistake to try to define cause in terms of mechanism because that generally leads back to where we started. What's a mechanism? A dictionary might respond that a mechanism is a process by which something is done or comes into being. But this seems just an oblique way of saying that a mechanism is something that turns causes into effects, and that's exactly where we started. I have yet to see a definition of mechanism that doesn't at least implicitly use the notion of cause. The fact that we can't define "cause" and "mechanism" without reference to one another suggests that they're closely connected. I'll talk about mechanisms as any kind of process that takes causes and produces effects.

So causal relations relate entities that exist in and therefore are bounded in time. I will refer to such entities as *events* or *classes of events*, because the word "event" suggests the transient character of causes and effects. On closer inspection, all the examples I noted have this character. For example, drugs don't cause addiction per se; rather, the class of event "drug consumption" causes the psychological and physiological events associated with addiction. Guns themselves don't cause death; the event of a gun being shot can cause the event that a living thing dies. Sometimes we talk in a non-temporally bounded way as if abstract properties can be causes and effects (increases in pressure cause increases in temperature; love causes beauty). But even here, the actual causes and effects as they manifest themselves in the world are physical entities that obtain for periods of time.

This is what distinguishes causal relations from definitions. A definition identifies a word or phrase with a set of conditions in the world. If all the conditions are met, then the word or phrase applies (i.e., the conditions are *sufficient*). Likewise, if the word or phrase applies, then the conditions are met (the conditions are *necessary*). If an object is geometric and has three sides, then it is a triangle. And, if it is a triangle, then it is geometric and has three sides. Causal relations don't associate linguistic entities with conditions. They associate events with other events.[1]

Experiments Versus Observations

The first thing an experimental psychologist learns is to distinguish an experiment from a correlational study (actually, this is the second thing they learn; the first is that theory without data is like cake without flour: all sugar, no substance). A correlational study

involves mere observation, what goes with what, when, and where (no why involved). Two attributes or variables are correlated if a value on one tends to be associated with a value on the other. Hat size is correlated with shoe size, the time on my watch is correlated with the time on your watch (assuming it's not broken), temperature is correlated with pressure. But the mantra of experimental psychology is that correlation is not causation. The fact that these things are correlated does not tell us *why* they are correlated. It does not tell us what generating mechanisms produced these correlations. To find out, we have to run an experiment.

Francis Bacon (1561–1626) was one of the earliest spokesmen for the advantages of the experimental method, when performed correctly: "For the subtlety of experiments is far greater than that of the sense itself, even when assisted by exquisite instruments—such experiments, I mean, as are skillfully and artificially devised for the express purpose of determining the point in question." Bacon thought an experiment was akin to torturing nature for its secrets: "For like as a man's disposition is never well known or proved till he be crossed, nor Proteus ever changed shapes till he was straitened and held fast, so nature exhibits herself more clearly under the trials and vexations of art than when left to herself."[2]

An experiment requires manipulation. Some variable, some potential cause (often called an independent variable), is chosen along with another variable, a potential effect (often called a dependent variable). The cause is then manipulated by setting it to two or more values and the effect is measured. If the value of the effect differs for different values of the cause, then we can infer that the cause has some influence on the effect; a causal relation exists. To illustrate, say you want to know if punishing children makes them more obedient. Then you need to vary punishment, the independent variable, by punishing some children more and other children less and measure their level of obedience, the dependent variable. If the children punished more are more obedient than the children punished less, and if the difference is big enough, you can conclude that punishment increases obedience. And if the children punished less are more obedient, then you have some explaining to do.

Much of the art of experimental methodology involves satisfying two requirements. First, the cause must be manipulated so that only the target cause is manipulated and not other potential causes accidentally. When manipulating punishment, you have to be careful not to also manipulate how much warmth you show the child. Second, the methodology must be sufficiently powerful to

allow a reasonable assessment whether a difference in the value of the dependent variable is real or just the reflection of random variability.

I'll say more later about what experiments allow us to learn that correlational studies don't. For now, the important point is that conclusions are drawn from experiments through comparison; the value of an effect must be compared in two different worlds: the world in which the cause is at one value and one in which the cause is at a different value, because a causal statement is a claim about two (or more) worlds simultaneously.

Causal Relations Imply Certain Counterfactuals

To say that A caused B seems to mean something like the following: A and B both occurred, but if event A had not occurred (and B had no other sufficient cause), B would not have occurred either. The critical point is that a causal relation doesn't merely imply that events happened together but that there's some generating mechanism that produces an event of one type when engaged by an event of another type. So in some other world in which the mechanism had not been engaged by the cause, the effect would not have resulted. This is what distinguishes a causal relation from a mere correlation. A correlation between two event types means that they move together; when one happens, the other tends to happen, too. A causal relation has the further requirement of *counterfactual* dependence.[3]

A counterfactual is a statement that concerns another possible world.[4] ''If only'' statements are good examples of counterfactuals (''If only I had a million dollars.'' ''If only my true love wasn't so tall.'' ''If only I hadn't eaten so much last night.''). All of these open thought or discussion about some other world, different from this one. Some counterfactuals are introduced by ''almost'' to bring into focus possible outcomes that didn't actually occur: ''The Giants almost lost yesterday.'' ''My lottery number almost came up.'' ''I almost won the Nobel prize (unfortunately, my discovery was 150 years too late).'' Counterfactuals actually appear in a huge variety of linguistic forms (''He could do it if he wanted to,'' ''I would have been happy if I hadn't taken that job''), sometimes with no special markers (''It's the things you don't say that you regret the most.'') Some counterfactuals aren't necessarily about other possible worlds but could be about this one (''She might have gone to the opera, but I think she went to the nightclub'').

A causal relation assumes a counterfactual, one that is related to *the effect would not have occurred if the cause had not*. This is a counterfactual because in fact the cause did occur and so did the effect (or, at least, they may well have). But what makes the relation causal is that the cause is responsible for the effect in that (again) if it hadn't occurred, the effect would not have either. Correlations make no such counterfactual claim. Shoe size and hat size are correlated, but that doesn't mean that if my feet swell, my head will swell too. Correlations are just about what goes together in the actual world. They're not about how things would be if the world were different.

For the same reason, counterfactual dependence also distinguishes a causal relation from an "association," a favorite term of psychologists to talk about how events elicit thoughts and behaviors. An association between two events is normally assumed to be formed by an organism who perceives that the events are correlated.

In sum, causal relations are more than descriptions of the way things are, the state of the world. They are processes that generate possibilities, whether those possibilities exist—whether they are actual—or counterfactual. Causal relations tell you how you get there from here, whether you're actually going there or not. As generative processes, they differ fundamentally from correlations, for a correlation is merely a description of what's been observed. So the first (or second) law of experimental psychology holds: correlation is not causation.

The great Scottish philosopher David Hume taught that causation cannot even be inferred from correlation.[5] On logical grounds, no number of observations permits a causal inference. No matter how many times A and B occur together, mere co-occurrence cannot reveal whether A causes B, or B causes A, or something else causes both. But one of the million dollar questions today (actually, research funding agencies have spent well over a million dollars on it) is whether Hume was right. Some recent theories of causality specify how causal inferences can be drawn from correlational data in certain cases. I'll touch on that question later. It's worth mentioning, though, that Hume also taught that people make causal inferences in everyday life anyway, despite their lack of justification in doing so. This kind of unjustified causal attribution is all around us. When a new administration enters government and the price of gas rises, there's a strong tendency to blame the new administration. In this kind of case, very different causal attributions may have just as much support: the forces that led to a rise in gas prices

may also have created the conditions for the new administration to be elected.

Enabling, Disabling, Directly Responsible: Everything's a Cause

So far I have pointed out that a causal relation implies a counterfactual. This isn't really saying much because people are not necessarily aware of the implied counterfactual; they may not know about it at any level. If you admit to chopping down the cherry tree, that doesn't necessarily mean that you're thinking, consciously or unconsciously, that the cherry tree could still be alive. Moreover, the relation between a cause and a counterfactual isn't simple, for it could depend on a host of other events.

Let's focus on a specific event to note how else it might have been and how all the other events responsible for that event (the *parent set* of that event) together determine how it might have been. Take your birth, for example. The events that together caused your birth include some actions by your mother, some actions by your father, the presence of food and oxygen in your mother's environment after conception, a favorable gestation, a safe place for birth, and so on:

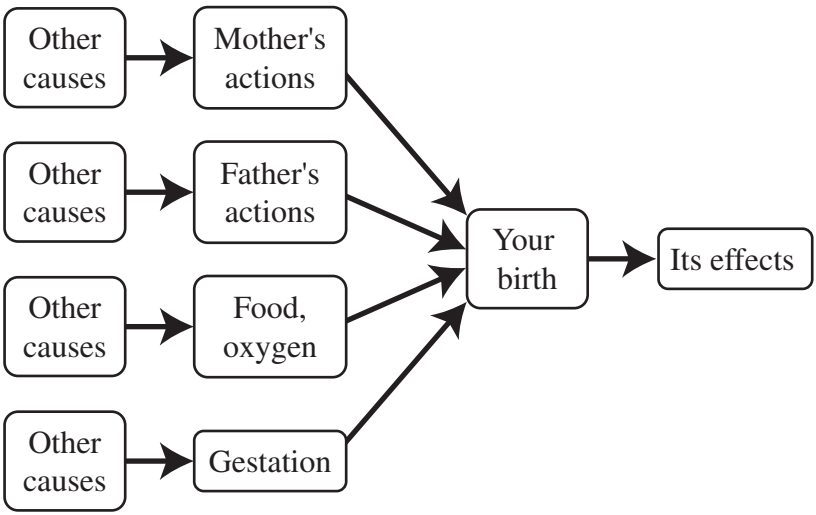

Figure 3.1

In an important sense, your mother's and father's actions are the ''true'' causes—they are at least the responsible people—while food, oxygen, and gestation are just ''enablers'' because they make the

critical event (your birth) possible or set the stage for it. You can also imagine "disablers" that prevent an event, like a chemical exposure that causes a miscarriage. But such distinctions can become fine legal points. Is birth control a disabler or a true cause (of not giving birth)? What's clear and what follows without any legal or moral analysis is that birth would not have happened without *all* the precursors. The event might well have been different (counterfactually) if any of those parental events had been different. Who knows how you would have turned out if gestation had been different or if your mother had enjoyed a different diet? Therefore, for simplicity, we'll use the word "cause" in a general way, to refer to all precursors, all variables that would have led to a different effect had they been different. The notion of enabling and disabling conditions is useful, and I'll use it at various points. But at the most basic level, where we just want to know about the mechanisms that make events happen, it is very useful to divide the world up simply into causes and effects.

Causes don't always act in concert to produce events. Sometimes an event has alternative causes, each itself sufficient for the effect. Death can be caused by trauma, uncontrolled growth, infection, and so on. In rare cases, more than one can conspire to cause death together. The arrow-drawings or graphs like figure 3.1 that I'll use in this book don't mark the difference between cases where causes act together (conjunctively) or separately (disjunctively) to produce an effect. But later I'll make this distinction by representing the way causes produce effects as simple mathematical equations (what are known in the trade as structural equations).

Problems, Problems

To see some of the problems in defining cause, consider the following story.[6] Betty throws a rock at a bottle. Charlie throws a rock at the same bottle. Both of them have excellent aim and can break the bottle almost every time. Betty's rock gets to the bottle first and breaks it. Charlie's rock would have broken it, but it got there too late. Did Betty's rock cause the bottle to break? Of course. Did Charlie's? Of course not. Yet, under the counterfactual test, if Betty hadn't thrown her rock, the bottle still would have broken, just like the bottle would have broken even if Charlie hadn't thrown his rock. In terms of the events, Betty's and Charlie's situations are the same. Both threw rocks and a bottle broke. And counterfactually they are the same. Yet they're different: Betty caused the bottle to break;

Charlie didn't. What's the difference? Somehow it seems important that Betty's rock touched the bottle, whereas Charlie's didn't. We can spell out a detailed causal chain for Betty, not for Charlie.

If we asked a different question, who's more to blame for the bottle breaking, then the difference between Betty and Charlie seems smaller. They'd both better run if the owner of the bottle shows up.

A variant of this story concerns a sheik whose wife wants to kill him because she just found out about his mistress. The mistress wants to kill him anyway to steal his money. The sheik's going on a long journey through the desert, so the wife puts poison in his water sack. Just before he leaves, the mistress puts a small hole in the water sack so that his water drips slowly out and he dies of thirst in the desert. The wife then reports the mistress, who gets thrown in jail for murder. There's no evidence against the wife. But is she guilty?

These problems disturb philosophers because they undermine the very definition of a causal relation. I have defined A caused B in terms of counterfactuals; I said that A and B occurred and if A hadn't occurred, B wouldn't have occurred either. In our story the mistress connives to let the sheik die of thirst and he does die of thirst. She would seem to be the cause of his death. And the wife tries to kill him with poison but fails. So she doesn't seem to be the cause of his death (she may be guilty of attempted murder, but her failure would presumably block a verdict of guilty of murder).

Yet both women are in the same position with regard to the definition. Both acted; both had the desired effect achieved. The sheik is dead. However, and this is the critical point, *both would have achieved the desired effect even if they had not acted*. The counterfactual part of my definition of cause was that the effect would not occur in the absence of the cause. But in the story, the effect would occur (the sheik would be dead) if the wife had not acted (because the mistress would have killed him). So according to this definition, the wife is not the cause of death. Similarly, the sheik would be dead if the mistress had not acted (because the wife would have killed him). So my definition suggests that the mistress is not the killer either. The counterfactual definition suggests that neither is guilty, that neither is the cause of death because neither's action changed what would have happened anyway. There must be something wrong with the definition.

Any definition of cause must be more elaborate to account for this sort of complication (and others). The definition that I gave earlier actually was slightly more elaborate. It said that the counterfactual part matters only if event B were to have no other

sufficient causes at the moment. But both women do have another sufficient cause of death. If the mistress fails, the wife will kill him, and if the wife fails, the mistress will kill him. If we were to rule out these alternative sufficient causes, if we ignored the wife when considering the mistress and ignored the mistress when considering the wife, then their respective actions indeed would be responsible for the effect. If we ignored the wife's poison, then the mistress's hole would be the difference between life (the counterfactual world) and death (the actual world), and vice versa for the wife. So this second definition that rules out alternative causes posits that both are, in fact, the causes of death.

We just can't get it right; both definitions fail to tell us that the mistress is the cause and the wife isn't. What's wrong with this definition of cause? At one level, it fails to be sensitive to the particular fine-grained causal pathway that takes us from action to effect. We can point to a chain of true causal links that take us from the mistress's action to the sheik's death (the hole in the canteen, the slow emptying of its contents on the desert floor, the sheik's look of surprise and dismay as he lifts the canteen to his lips, his gradual weakening from dehydration . . .). We can do no such thing for the wife because the sheik didn't die of poison. But in the end this doesn't solve the problem because of the critical words "causal pathway" in the description of the chain. I'm trying to define *causation,* but it seems like cheating to do so by appealing to a *causal* pathway. That's like the left hand paying a debt to the right hand. Maybe it's possible to describe this pathway without using the notion of cause by appealing to specific laws. Physical laws govern how water drips out of a canteen, for instance, and if we can describe the whole chain in terms of such well-known laws with independent motivation, we might be able to define cause without referring to cause.

Another possibility is to make the definition more sophisticated. This is what the legal philosopher J. L. Mackie did.[7] I'll present his idea briefly to offer a taste of it even though it's going to be a bit of a tangent. To really understand it, I recommend reading the original. Mackie pointed out that what we call a cause is actually a part of a larger entity. Whenever an effect is caused, a large number of true conditions in the world are responsible for it. If a rock breaks a bottle, it's because (1) the rock was heavy enough (but not so heavy that it couldn't be thrown); (2) the rock was moving at sufficient speed; (3) the bottle had sufficiently weak bonds holding it together; and so forth. Altogether, these conditions are sufficient for the effect, and anything we call a cause is part of such sufficient

conditions, indeed a necessary part because we wouldn't call it a cause if the effect would have happened even if the condition referred to wasn't there. Of course, the set of conditions sufficient for the effect aren't together necessary, because there may well be some other set of sufficient conditions that would lead to the same effect. So the condition at issue is a cause if it is Insufficient itself for the effect, but it is a Necessary part of an Unnecessary but Sufficient set of conditions. That's a mouthful both linguistically and conceptually. So Mackie called it INUS for short. A condition is a cause if it is an INUS condition. This reduces the linguistic problem, but it still leaves a conceptual mouthful.

To see how INUS solves our problem, think about the wife's poisoned water and the mistress's hole in the canteen. The hole in the canteen is one (I) of the critical elements (N) in one (U) of the long set of conditions that must have obtained for the sheik to die (S). So it was a cause of death. The poisoned water, in contrast, is not a critical element of a sufficient set of conditions because the set of conditions that include drinking poisoned water did not actually kill the sheik. It would have to include conditions like "the sheik drank the poisoned water," but it can't because such conditions didn't actually obtain. It's critical of INUS conditions that they refer to conditions that are actually true of the world. But when an event didn't actually lead to an effect, then something has to be not true of the world and the INUS condition won't be met.

The discussion in this section has been using the word "cause" in a different sense from the rest of this book. As I'll discuss at length in the next chapter, the rest of this book will use "cause" to mean anything that can be represented by an arrow in a causal graph. To foreshadow, both the wife's poisoning and the mistress's puncture can be represented as arrows in a causal graph:

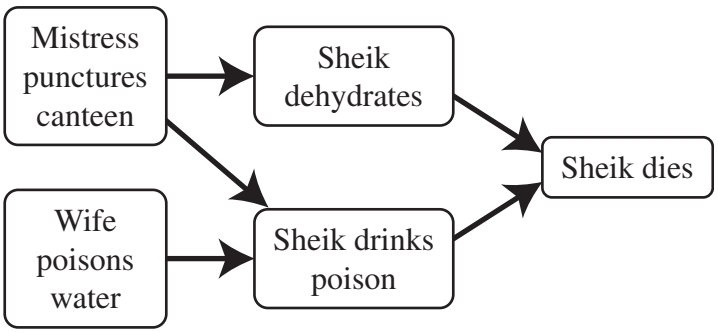

Figure 3.2

In that sense, all of these events (that may or may not occur) are causes. In this section, I have been discussing which event in a causal graph is the *actual* cause. Both the wife's and mistress's actions may be causes in the sense of being relevant in the graph, yet only the mistress's action (I think we agree) is the *actual* cause of the sheik's death.

The graph actually offers a simple explanation. There's a link from the mistress's puncture to the sheik drinking poison because the puncture *prevented* the sheik from drinking the poison (due to the puncture, there was nothing for him to drink). The puncture blocked the chain of events from the wife's poison to the sheik's death from coming to pass. There's no such preventive link from the wife's poisoning to the chain of events from the mistress's puncture to the sheik's death. So the reason that the mistress is at fault is because her action was not prevented from having its intended effect; the wife's action was prevented from having its effect.

Several philosophers have suggested that actual causes can be read off a causal graph using this kind of logic.[8] Very roughly, an event is an actual cause of another event if both of the following are true:

1. Both events actually occurred.
2. If the first event had not occurred (counterfactually), then the second event wouldn't have either, even when all events not on the causal path from the first to the second are assumed to have occurred (as they actually did).

If we assume that the canteen was punctured and the sheik dehydrated, then no change to the amount of poison would prevent the sheik from dying. Hence, the wife's poisoning is not a cause. In contrast, if we assume that the wife poisoned the water but the sheik didn't drink the poison—what actually happened—then the mistress's puncture is the difference between life and death. So the mistress is guilty.

This kind of analysis doesn't provide a definition of cause that's independent of cause. After all, it depends on chains of arrows, and arrows are nothing but causes. What it does do though is offer an idea about how people go about determining what the actual cause of an event is from a bunch of causal knowledge.

Could It Be Otherwise?

A frequent problem with scientific theories, especially theories of how the mind works, is that they are too powerful. A theory that can explain everything, both what is true and what is false, in the

end explains nothing at all. A parody of such a theory would be one that attributes thought to a little person in the head (a homunculus). If the explanation of how I think is that another person is thinking for me, then we haven't gotten anywhere at all. This would just be the tip of an explanatory circle in which X is explained by Y, which is in turn explained by X.

But we have to be careful, because theories that have this essential character often come dressed in other clothes, so it's hard to identify their circularity. Consider a simple theory that attributes all learning and memory to associations. Learning consists in the construction of associations between things in the world or between actions and things in the world. (A theory like this was once made very popular by B. F. Skinner.) Notice just how powerful a theory it is. Whatever you tell me you've learned, I can describe the associations that you've constructed. If you've learned a language, then I can wave my hands about the associations between objects and words, and the associations between parts of speech like nouns and verbs that you must have acquired. But, of course, I could describe an infinite number of associations that you haven't learned, too. I never learned that "outside" means where clouds are, even though, where I come from, one usually sees clouds when one goes outside. So the real question is what's the difference between the things you have learned and the things you haven't. Simply listing them after we already know they exist, even in the form of a list of associations, doesn't help. Because the theory can explain everything, it explains nothing. Of course, sophisticated theorists are aware of this problem, so good associationist theories do say more. Generally, they describe the process of learning in some detail.

The problem of circularity is pervasive when people try to explain mental processes as the result of evolution. I, for one, am a great believer in an updated version of Charles Darwin's process of natural selection. One of the beauties of natural selection is that it is so simple and yet so powerful that it provides an explanation for the origin of all living things from the primordial soup to current, complex biological organisms. It is therefore tempting to use it to explain the most impressive biological entity around, the human mind. After all, presumably the mind did originate during a process of evolution. It may also have been the side effect of the evolution of other functions (my favorite is the idea that the mind developed alongside the brain, which evolved as a large-scale cooling system for the rest of the body). Nevertheless, there has got to be some evolutionary story about how the human mind came to have its

current form. The problem is that there is also an evolutionary story for how the mind did *not* turn out. The theory of evolution, because of its elegant simplicity, is powerful enough to allow the construction of a plausible story for any possible mind that didn't happen to evolve. The trick is merely to find some adaptive function that such a mind served and then dramatize why that function was so important in the era of greatest change to the species. Evolutionary theory is a powerful source of stories, so good evolutionists are careful when they go about storytelling, always making sure that they have enough facts to warrant a theory and that their theories stay close to those facts.

Does the notion of cause suffer from the same problem? Is it so powerful a concept that it can explain everything, even what is not observed?

Not All Invariance Is Causal

Not only is the notion of cause not general enough to explain all possible properties of the mind, even those that don't exist; it is not even so general that it can explain all properties that *do* exist. Some of the invariance that people are sensitive to is not causal at all.

People pay attention to part-whole relations, for example. We are experts in recognizing how nose, mouth, eyes, cheeks, compose a whole face or how abstract shapes compose an abstract pattern, like a plaid. We immediately perceive how the parts of a view—the people, objects, the background—relate to make up a scene. The human capacity to recognize and create patterns out of parts is masterly (pigeons are pretty good at it, too, it turns out). People are also sensitive to relations among classes (e.g., that a chicken is a fowl and a hammer a tool). Syntactic structures in language are not causal (although, as we'll see in chapter 11, they, too, encode some causal properties).

So people appreciate and use a variety of relations that are not causal. Many of these relations can be described using other kinds of logic, like set theory. Set theory is a well-developed branch of mathematics useful for talking about various kinds of relations. For instance, a part-whole relation is reminiscent of the relation between a subset and a set; a subset is a part of a set. A nose is a part of a face; the features that make up a nose are a subset of the features that make up a face. Set theory is also useful for thinking about relations among classes or categories. For example, according to the current American tax code, the class of trucks includes the subclass sports-utility vehicle (SUV). In other words, the set of SUVs is a

subset of trucks (according to the U.S. Internal Revenue Service anyway). But not all set-subset relations are the same. Part-whole relations—known as *partonomies*—differ from subclass-class relations—*taxonomies*—in that a subclass is a member of the class, whereas a part is not a member of the whole (an SUV is a truck, but a nose is not a face). Also, a subclass inherits all the properties of the class, whereas a part does not inherit the properties of the whole (everything that is true of a truck is by definition true of an SUV, but plenty of things are true of faces—like their ability to display emotions—that are not true of noses).

Sets support all kinds of logical relations. If A is a member of B, and B is a member of C, then A is a member of C. If A is true and B is true, then ''A and B'' is true. And people are highly sensitive to much of this structure. But set theory is severely limited in what it can easily represent. As we'll see, it is unable to represent the logic of causal intervention.

Probability theory also captures invariance. But much of the invariance it describes results from the operation of causal mechanisms. The reason that the probability of rolling double-six with two fair dice is 1/36 has to do with counting the possible causal outcomes of rolling the dice separately. The reason that the weather forecaster predicts rain with probability .7 generally has to do with a causal model that predicts future weather outcomes by extrapolating from current conditions. The reason that the conditional probability is high that the ground is wet, given that it has rained, is that the former is a causal effect of the latter. So the invariance of probabilities is really just causal invariance in disguise.

Admittedly, not all probability relations have a causal basis. Quantum effects in particle physics are probabilistic, and philosophers argue about whether or not they have a causal basis. But such effects are notoriously hard for people to think about, so they don't provide much of a challenge to a causal model of thought. Also, some probabilities reflect degree-of-belief, not causal mechanisms. My subjective probability that the ancient Sabines inhabited a powerful city state reflects confidence, not causality. But even my confidence is generated by causal mechanisms, like the ease of recalling paintings and stories of Sabines and causal explanations about why Sabines might seem familiar.

In general then, causal relations are not the only kind of invariance useful for representing the world. There are various kinds of mathematical representations, as well as logical and probabilistic representations. But noncausal forms of invariance are less useful

than causality for describing relations among events because they don't naturally describe the processes that generate those events and because, therefore, they fail to support key forms of counterfactual inference as directly as causal models do. In short, only causal models represent the invariance that tells us what the effects of our and others' actions would be. As a result, people seem to be particularly adept at representing and reasoning with causal structure.

4

Causal Models

Hopefully you now agree that, at least according to human perception, the world is full of causal systems composed of autonomous mechanisms that generate events as effects of other events. In this chapter, I will try to make this idea more precise and therefore more clear by making it more formal. I will introduce the causal model framework as an abstract language for *representing* causal systems. First, I will discuss it as a language for talking about causal systems. Later, I'll talk about what parts of it might have some psychological reality. These are not the same, of course. You might have a great system on your computer for bookkeeping, but the very reason you have it on your computer is because you don't have it built into your mind. Similarly, there are excellent systems for representing causal relations that might or might not describe how people think about causality. So let's start with a really good system for representing causality before we talk about how people do it.

The causal model framework was first spelled out in detail in a book published in 1993 by Spirtes, Glymour, and Scheines. The framework is based on a mathematical theory for representing probability called Bayesian networks. Many computer scientists, statisticians, and philosophers have contributed to its development, but I will rely to a large extent on the very complete treatment offered by Judea Pearl in *Causality*, published in 2000. The framework is a type of graphical probabilistic model. It's probabilistic in

that it allows uncertainty or ignorance about whether an event will occur. It allows us to reason about events when we are unsure about what has happened, what will or would happen, and even about how events lead to one another. All we have to know is how likely events are and how likely they are to be caused by one another. In particular, causes don't always have to produce their effects; they only have to produce them sometimes.

The framework doesn't insist on probabilistic relations, however; if a cause always produces an effect, that is, if the cause and effect are related deterministically, that's all right, too. The framework is graphical in that a graph composed of nodes and links (like the graph of your birth in the previous chapter) represents the causal structure of a system, with nodes corresponding to events or variables in the causal system, and directed links (arrows) between nodes corresponding to causal relations.

The Three Parts of a Causal Model

In this kind of scheme, three entities are involved in the representation: the causal system in the world (i.e., the system being represented), the probability distribution that describes how likely events are to happen and how likely they are to occur with other events—how certain we can be about each event and combination of events—and a graph that depicts the causal relations in the system (see fig. 4.1).

In this picture, a link from one entity to another means the second *represents* the first. The probability distribution and graph both represent the world, and the graph also represents the probability distribution. Representation is a key concept here. In this context, a set of symbols represents a set of elements if the symbols are related to one another in the same way that the elements being represented are related to one another. A highway map is a representation of a region if the distances between places on the map are proportional to distances between corresponding places in the region. A caricature of Abraham Lincoln is a representation of Lincoln because the parts of the caricature's face are related to one another in the same way that the parts of Lincoln's own face were related (though the caricature's relations might be exaggerated). A representation has to be simpler than the thing being represented because it abstracts from the thing being represented; it mirrors some aspect of it. A representation can't mirror all aspects; if it did, it wouldn't be a representation, it would be the thing itself.[1]

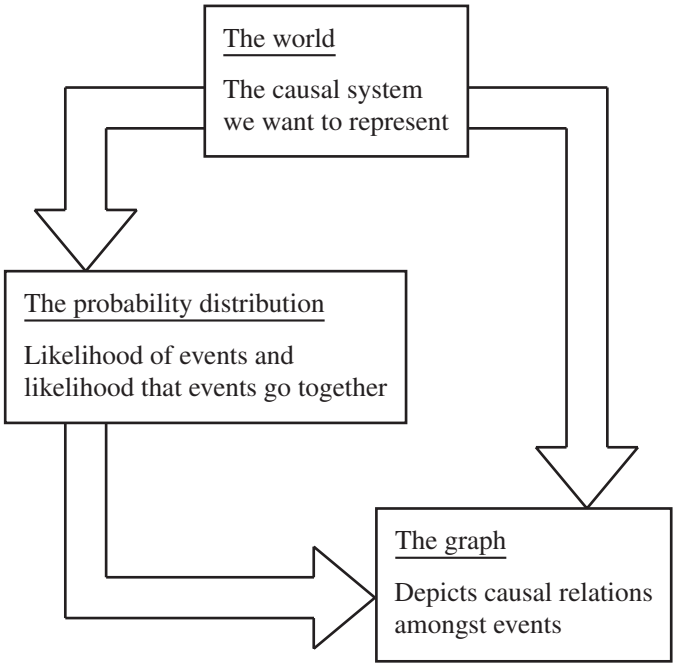

Figure 4.1

Probability distributions are representations of the world because they capture certain relations in the world. Namely, they specify how much confidence we can have that an event will occur or that an event will occur if we know that another one has.[2] They are also simpler than the world because they don't specify everything about it. They don't specify, for example, how good an event is, only how much confidence we can have in its occurrence. Graphs are representations of probability distributions because they specify the causal relations among events that are implicit in the probabilities. In other words, they depict the causal relations responsible for the probabilistic ones. They are simpler than probability distributions because they don't show every probability; rather, they show only the structure of the causal mechanism that generates the probability distribution.

Let's illustrate with fire, a causal system in the world that we're all familiar with that relates sparks, oxygen, an energy source to feed it, as well as other things, as shown in figure 4.2.

The probability distribution consists of a set of marginal probabilities and conditional probabilities. A marginal probability is the raw probability of an event like the probability of fire, written P(fire),

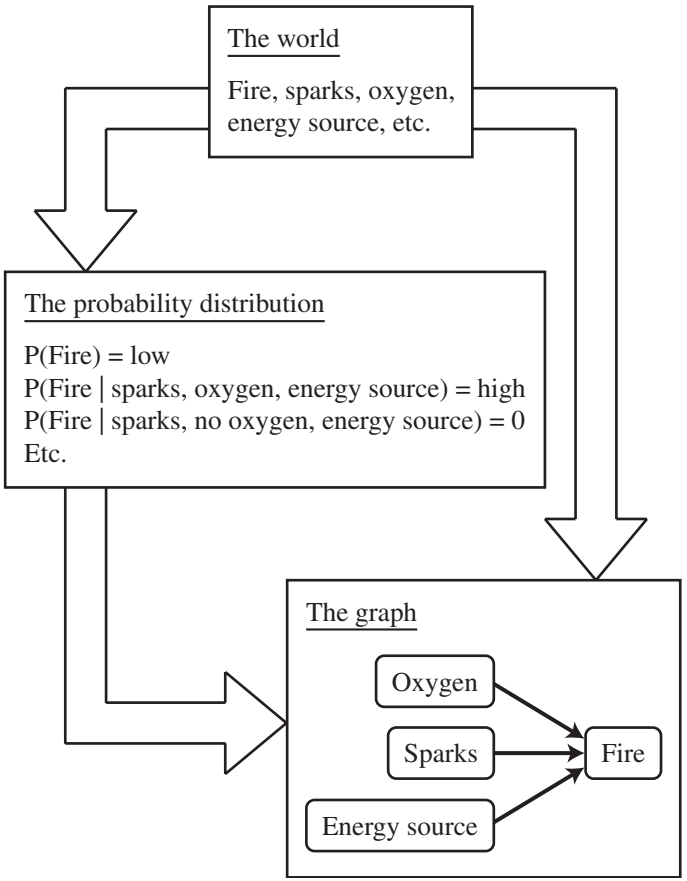

Figure 4.2

or the probability of sparks, P(sparks). A conditional probability is the probability of an event once you know the value of one or more other events. If you know there are sparks and oxygen and an energy source around, then you might ask what the probability of fire is, given this knowledge, conventionally written as P(fire | sparks, oxygen, energy source), where the symbol | means ''given.'' Some people believe that all probabilities are conditional probabilities because all probabilities are assigned based on some knowledge.

Independence

The graph represents the causal relations in the system, and it constrains what the probability distribution can look like. For instance,

if there's no path between two events in the graph, then this should be reflected in the probability distribution. For instance, music is unrelated to fire, so if the presence of music were included in the causal system, it would be represented as a disconnected node in the graph:

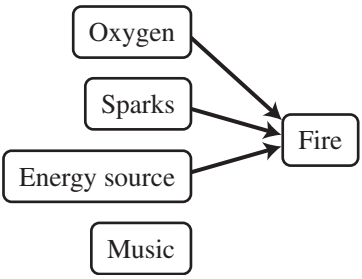

Figure 4.3

The probability distribution would reflect this by making sure that

$$P(\text{fire}) = P(\text{fire} \mid \text{music}) = P(\text{fire} \mid \text{no music}).$$

In English, the marginal probability of fire is the same as the conditional probability of fire given music and also the same as the conditional probability of fire given no music. This relation between fire and music is called independence because the probability of fire is the same whether music is playing or not.

In sum, a causal model representation has three aspects: the world being represented, an algebraic representation of it in terms of probabilities, and a graphical representation of the causes that generate the probabilities. One of the most important aspects of the world captured in both the probabilistic and graphical representations is independence, when events have nothing to do with each other. Every time we can say events are independent, we have said something important because it simplifies the graph by removing links and makes it easier to do calculations with probabilities. The most significant qualitative information in a causal model is that two events are independent, because independence lets us simplify, and simplification while maintaining veridicality (accuracy) is the key to an effective representation.

Structural Equations

It is easiest to think about the graph and the probabilities as parts of a single representation. You can do so by associating each node of a causal model with a set of probabilities or conditional probabilities.

Each node is the joint effect of all links pointing into it. Fire is the joint effect of sparks, oxygen, an energy source, and so on. So the effect (fire) is fully described by stating the probability of fire under all possible combinations of its causes:

P(Fire | sparks, oxygen, energy source) = high
P(Fire | sparks, oxygen, no energy source) = 0
P(Fire | sparks, no oxygen, energy source) = 0
P(Fire | sparks, no oxygen, no energy source) = 0
P(Fire | no sparks, oxygen, energy source) = very low
P(Fire | no sparks, oxygen, no energy source) = 0
P(Fire | no sparks, no oxygen, energy source) = 0
P(Fire | no sparks, no oxygen, no energy source) = 0

To keep the variables simple, I'll assume that each takes on one of only two possible values—present or absent—even though in reality each variable is essentially continuous (the fire could be any size and there could be various quantities of sparks or of oxygen, etc.). Even with this simplification, $2^3 = 8$ conditional probabilities (all possible combinations of the presence or absence of sparks, oxygen, and energy source) would be required to describe this one little mechanism producing fire. To describe the causal system fully, we'd also need to know the marginal probabilities of sparks, of oxygen, and of an energy source. So a lot of numbers are necessary ($8 + 3 = 11$). But these 11 numbers would be a complete representation of this common causal system.

The goal of structural equation modeling is to make the representation smaller and more elegant by specifying how the causes combine to produce the effect. Causes can come in many forms. Some are direct causes, others enabling or disabling conditions. Multiple causes might produce an effect jointly or individually. They might add up or multiply to produce the effect.

Structural equation modeling expresses the precise nature of the functional relation representing the mechanism determining each effect using a different representational scheme than probabilities and graphs, but one that carries the same information (technically speaking, an isomorphic representation). Each effect is associated with a mathematical function expressing exactly how it is produced by its causes. For example, if a spark, oxygen, and a source of energy are required to produce fire, the following equation could represent a causal mechanism:

Fire = f(spark, oxygen, energy source)

where f means conjunction—that all causes are required. The way I've written this equation ignores the fact that the function is

actually probabilistic. A more proper equation would include an additional variable that would be called error or noise. It would represent the contribution of sources of randomness that make the relation probabilistic.

This equation is a more compact and simple way of writing what is expressed by the long list of conditional probabilities shown. Instead of describing the mechanism as a list of the probabilities of the effect for each possible combination of causes, it describes the mechanism more directly by expressing how the causes are combined to produce the effect. Causal graphs are good for expressing a complex system of causal relations; structural equations are good for expressing the specific function relating a set of causes to their effect.

What Does It Mean to Say Causal Relations Are Probabilistic?

A causal relation is probabilistic or is affected by random factors if the combination of known causes isn't *perfectly* predictive of the effect. Some combination of causes might usually produce the effect but not always; conversely, the effect might usually be absent in the presence of certain causes yet occur sometimes nevertheless.

Sources of randomness can sometimes be reduced to other causal mechanisms that we happen to be ignoring. The wind is an important factor in whether a fire occurs, but it makes sense to consider the causal mechanism of a spark mixing with oxygen and fuel to understand the physics of fires and to understand the key controllable ingredients. So we might ignore the wind and treat it as a random factor. And there could be other random factors, such as how dry the environment is and the ambient temperature. Sometimes our fire doesn't start even though all the critical elements are there, but instead of causally explaining why not, we just attribute it to one of the various random factors. These factors are random only in the sense that we're ignoring them by bundling their effects into a single variable that we call random that makes the functional relation between causes and effects not perfectly predictable but rather probabilistic.

Some theorists believe that some types of probability are not due to what we ignore, but rather are due to intrinsic unpredictability in certain causal mechanisms. For instance, some physicists consider some phenomena of quantum physics to be intrinsically probabilistic. We won't have to worry too much about this possibility, though. For the kinds of medium-scale causal systems that

people tend to think about in everyday life, randomness is produced by what we ignore, not by the fundamental nature of events.

Causal Structure Produces a Probabilistic World: Screening Off

One of the central ideas of the causal modeling framework is that stable probabilistic relations between the observed variables of a system are generated by an underlying causal structure. In other words, the world we see around us with all its uncertainty can be attributed to the operation of a big, complicated network of causal mechanisms. On a smaller scale, particular kinds of causal structure will lead to particular patterns of probability in the form of particular patterns of dependence and independence.

Let's see how direct causal relations depicted by arrows in a graph correspond to relations of dependence and independence. In the simplest case, if we ignore the direction of arrows in a causal graph and see that there's no route from one variable to another, if the two variables are disconnected, those two variables should be unrelated, or probabilistically independent. In contrast, if a causal arrow points from one variable to another, they should be related, or dependent. In the case of a *causal chain*:

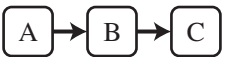

Figure 4.4

A and C should be dependent because A is an indirect cause of C. Imagine that A is lightning, B is fire, and C is heat.

Figure 4.5

On this model, wind and heat are dependent because you're more likely to have heat if you have lightning than if you don't have lightning. Changes in A should produce changes in C. But if we hold B fixed, A and C should be independent because the effect of A on C is mediated by B. In more static terms, A tells us nothing about C when we already know the value of B. In particular, if we know there's a fire of a certain size, then we learn nothing about the presence of heat by learning whether there's lightning. Similarly, if we know there isn't a fire, knowing about lightning doesn't tell us

anything additional about heat. A and C may be dependent, but they are *conditionally independent given B*. This is one example of what mathematicians call the Markov condition, or what psychologists call the "screening-off property" of causal graphs. B screens off A from C by virtue of mediating A's causal effect on C.

Screening off also arises with *causal forks*:

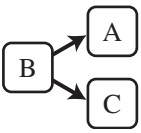

Figure 4.6

In this case, B is a common cause of both A and C. For example, let A be damage and B and C again be fire and heat, respectively. Because B causes both, A and C are again dependent (they tend to occur when B does and not to occur when B doesn't). So the relation between A and C is again mediated by B. As a result, if we fix B, A and C are no longer related; they are independent conditional on B. If we already know there's fire, or if we know there's no fire, learning there's been damage tells us nothing new about the presence of heat. So the conditions for screening off are also met with a fork: A and C are dependent but independent conditional on B. B screens A off from C.

The third basic kind of relation among three variables is known as a collider or *inverted fork*:

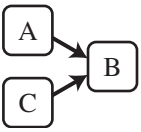

Figure 4.7

In this case, B is a common effect of both A and C. Let A be lightning and C a smoldering cigarette and B again is fire. In this structure, A and C are independent if the value of B is unknown because they have no common causes and they don't cause each other. The probability of lightning doesn't change if there's a smoldering cigarette on the ground. But if B is known, then they become conditionally dependent. If we know there's a fire, then the probability of lightning decreases if we learn there was a smoldering cigarette because the cigarette explains the fire. The fire provides less evidence

for lightning because the fire is already explained. This is called explaining away and will be discussed in chapter 6 in the context of discounting.

Equivalent Causal Models

I just claimed that observed probabilities are generated by causal models and that therefore knowledge of causal structure allows us to make inferences about dependence and independence. Conversely, given the observation of certain dependencies and conditional dependencies in the world, one can infer something about the underlying causal structure. Say you measured A and B at various points in time and you noticed they changed together: when one is high, the other tends to be high; when one is low the other tends to be low (in statistical terms, they are correlated). In other words, they are dependent. This could occur for many reasons. A could be the cause of B,

Figure 4.8

or B could be the cause of A,

Figure 4.9

or A could be an indirect cause of B,

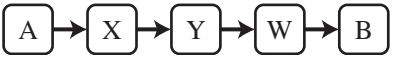

Figure 4.10

or both could be caused by some other variable V,

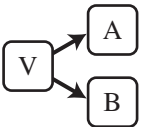

Figure 4.11

Indeed, you could produce many other causal models explicating why A and B are dependent. Some of these possibilities could be tested by examining other dependencies. For instance, if A turned out to be independent of X, then the chain in which A is an indirect cause of B could be ruled out.

Let's consider the possibilities with only three variables, A, B, and C. If you know that A and B are dependent and also that A and C are dependent but independent given all possible values of B, then A and C must be causally related and B must screen off A from C. Only three causal models that relate A, B, and C to one another are possible to express these relations.[3] The first two are chains and the third is a fork.

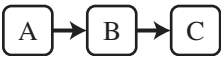

Figure 4.12

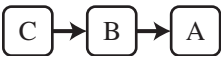

Figure 4.13

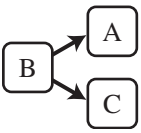

Figure 4.14

These are called Markov-equivalent (this is different from the Markov condition!) because they cannot be distinguished using relations of independence/dependence or indeed any kind of mere probabilistic information alone. Some other kind of information is required, like the temporal order in which the three events occur or the effect of an intervention, as I'll discuss in the next chapter.

Inferring Causal Structure Is a Matter of Faith

Making inferences from dependencies in the world to causal structure requires two assumptions concerning the relation between a causal graph and the probabilistic patterns of data it generates. Recall that the parent of a variable X is the set of all variables that feed into X, all its causes considered together. The first assumption says that a variable's parents screen it off from the parents' parents (its grandparents and other ancestors, if you like). This

assumption, known as the *causal Markov condition*, suggests that the direct causes of a variable—its parents—render it probabilistically independent of any other variables except its effects. This a very useful assumption, because in practice it means that we can determine the value of a variable by examining the value of the variable's parents along with the values of its effects, if we happen to know them. We don't have to go back through chains of indirect causes. This condition will hold so long as a causal graph explicitly represents any variable that is a cause of two or more variables in the graph.

The second assumption, the *stability assumption* (sometimes called *faithfulness*), stipulates that any probabilistic independencies in the data should arise solely because of causal structure and not because of mere chance. Imagine we flip two coins A and B. We win $10 if they come up the same (both heads or both tails), and we lose $10 if they come up differently. A and B are obviously independent events: P(A is heads|B is heads) = P(A is heads) = .5. Paradoxically, winning is also apparently independent of both A and B: P(WIN|A is heads) = P(WIN|A is tails) = P(WIN|B is heads) = P(WIN|B is tails) = P(WIN) = .5. That is, whether A comes up heads or tails, the probability of winning is .5; similarly for B. Moreover, the overall probability of winning is also .5. The likelihood of winning apparently has nothing to do with the outcome of flipping either A or B, yet, causally speaking, winning is completely determined by the outcomes of A and B:

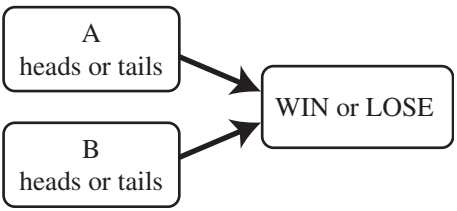

Figure 4.15

Notice, though, that the apparent independence disappears if the probability of A and B is anything other than .5. Imagine that A and B are biased; they have bits of heavy gum stuck onto one end such that they come up heads three out of four times. A and B are still independent: P(A is heads|B is heads) = P(A is heads|B is tails) = P(A is heads) = 3/4. But winning is no longer apparently independent. You win if both come up heads, which happens with probability $3/4 \times 3/4 = 9/16$, or if both come up tails, which happens with probability 1/

$4 \times 1/4 = 1/16$. So the probability of winning is the sum of these possibilities: $P(WIN) = 9/16 + 1/16 = 10/16 = 5/8$. But if you know that A is heads, then you win only if B comes up heads, which happens with probability 3/4. So the probability of winning given A is heads is $P(WIN|A \text{ is heads}) = 3/4$, which is not the same as P(WIN). Now WIN is dependent on A (and B by similar reasoning) as it should be.

To conclude, the apparent independence between winning and the outcomes of the coins was limited to the very special case where $P(A \text{ is heads}) = P(B \text{ is heads}) = .5$. It's really just a coincidence of values, not a true relation of independence. The stability assumption is that *stable* independencies arise from the structure of a causal model, not from mere coincidence. The result of flipping coin A is independent of the result of flipping coin B in this sense because the coins have no causal influence on one another. Two variables are expected to be truly unrelated or independent in the absence of certain kinds of causal paths between them. In contrast, *unstable* independencies arise from some coincidental set of values of the conditional probabilities in the model. Situations can arise when two variables satisfy the definition of probabilistic independence yet are causally related, as in the example just described of winning if, and only if, two fair coins both take on the same value. Such situations are unstable in the sense that the apparent independence disappears as soon as one of the relevant probabilities changes. In the example, winning was no longer independent of the coins once we imagined that the coins were no longer fair but instead had a probability of heads different from .5.

By assuming that unusual coincidences of value of this kind don't arise very often in the world and that they therefore can be ignored, we can use independence to help us decide on the true causal structure among variables. We make this stability assumption all the time in real life. For instance, we generally assume that if we see one person, then we should assume that there is only one person present, not that there's another person behind the first in exactly the position that would obscure the second person. The probability of two (or three or four . . .) people when all you see is one is so low that it should be neglected. The causal model analog is that we should assume that the independencies we observe arise from actual causal mechanisms and not from highly unlikely coincidences. The assumption derives from the desire to pick out stable mechanisms rather than transitory events.

In sum, patterns of probabilistic relations, dependence and independence, imply certain causal structures; from relations of

independence we can infer something about the underlying causal system. To do so, we have to make two reasonable assumptions about how the world works. By assuming screening off, that indirect causes (ancestors) have no effect when parents' values are fixed, and stability, that independence is not coincidental but arises only in the absence of causal relatedness, we can infer a lot about causal structure by observing the world and in some cases can identify it uniquely. Sometimes we cannot identify a unique causal structure, but we can still limit the set of possible causal structures to a Markov-equivalent set: those causal structures consistent with our observations.

The Technical Advantage: How to Use a Graph to Simplify Probabilities

The reason that statisticians and people interested in artificial intelligence care about all the fine details of graphs and probability distributions is that graphs can be used to simplify calculations; sometimes they help so much that they make calculations possible that would otherwise be just too hard. By representing what variables are independent of one another, graphs tell us what we can ignore. Often, much can be ignored.

I'll now try to explain how graphs help, although I won't do so with mathematical precision, and I won't even try to justify the claims (I'll leave that to more technical expositions). The following discussion is a little technical, as technical as this book will get, and you can skip it without losing the thrust of the rest of the book. But it's worth looking at, if only briefly, as it's the mathematical heart and soul of graphical probability models. It's why theorists bother with this kind of method.

Consider a small causal system composed of eight binary variables (A, B, C, D, E, F, G, H: each takes two values, say **on** and **off**). A complete probabilistic description of this system is called a joint probability distribution and is written as follows:

$$P(A, B, C, D, E, F, G, H).$$

To fully specify the joint probability distribution requires that we state the probability of all possible combinations of the variables' values):

$$P(A = \text{on}, B = \text{on}, C = \text{on}, D = \text{on}, E = \text{on}, F = \text{on}, G = \text{on}, H = \text{on}) = p_1,$$
$$P(A = \text{off}, B = \text{on}, C = \text{on}, D = \text{on}, E = \text{on}, F = \text{on}, G = \text{on}, H = \text{on}) = p_2,$$
$$P(A = \text{on}, B = \text{off}, C = \text{on}, D = \text{on}, E = \text{on}, F = \text{on}, G = \text{on}, H = \text{on}) = p_3,$$
etc.

If we spelled out all individual probabilities, we'd end with $2^8 - 1 = 255$ p values. (2^8 because there are 8 variables with 2 values each. We subtract 1 because the sum of all the ps must equal 1.0; the probability that the system is in one of the 256 states is 1.0. Therefore, we can figure out the last p by subtracting all the other ps from 1.0.) Each p is a parameter of the system so we see that a full description requires 255 parameters.

But if we know something about the structure of the causal graph that relates the variables, we can use that knowledge to reduce the number of parameters needed to describe the system. Imagine that the variables are related in the following way:

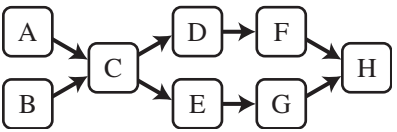

Figure 4.16

This graph shows a lot of independence relations among these variables: A is independent of B, and A and D are conditionally independent given C, along with many other independence relations. These independence relations can be taken advantage of to rule out many of the joint probability distribution's parameters. The trick is to read the joint probability off the graph starting from the root nodes (the initial causes) and progressing through the causal graph by following links. The fact that variables' values depend only on their parents and not on their parents' parents allows a simple technique for writing the joint probability distribution in a simple form. In this case, it turns out that

$$P(A, B, C, D, E, F, G, H) = P(A)\cdot P(B)\cdot P(C|A,B)\cdot P(D|C)\cdot P(E|C)\cdot P(F|D)\cdot P(G|E)\cdot P(H|F,G).$$

Notice that we obtain the new form by conditioning each variable on its parents. This equation may look scary, but in fact it is a lot simpler than specifying the probabilities of each state individually. To describe it fully requires knowing P(A), a single number or parameter, P(B), another parameter, as well as the conditional probabilities. They require more parameters. P(C|A,B) requires four:

$$P(C = \text{on} \mid A = \text{on}, B = \text{on})$$
$$P(C = \text{on} \mid A = \text{on}, B = \text{off})$$
$$P(C = \text{on} \mid A = \text{off}, B = \text{on})$$
$$P(C = \text{on} \mid A = \text{off}, B = \text{off})$$

Again, we don't need the probability that C is off because that can be obtained by subtracting the probability that C is on from 1.0. P(D|C), P(E|C), P(F|D), and P(G|E) require two parameters each, and finally P(H|F,G) requires four. So, all told, we need 18 parameters once we take into account the structure displayed by the causal graph. That's an immense savings over the 255 required without using that structure. In many cases, the 255 parameters would be unobtainable. For instance, if we have to run an experiment to obtain each one, we'd have to run 255 experiments and we may not have time or money to do so. But if we only need 18 parameters, obtaining them may well be possible.

Some real-life causal systems have hundreds or thousands of variables. Think of all the relevant variables in a political system or in a modern airplane. The space shuttle has about a quarter of a million operational components. The more variables a system has, the more potential savings there are from parameter reduction if we have a causal graph that we can trust.

5

Observation Versus Action

The discussion so far has focused on observation, what our causal graphs suggest that we can expect to see in the world and what our observations of the world allow us to infer about causal structure. And probability talk seems a good way to talk about what we observe and what we should expect to observe. But there's more to causality than that. Causality doesn't just tell us what to expect as passive observers, but what to expect when we take action, when we act as agents and intervene on the world. Causality concerns the effects of the actions we take. It also concerns the effects that actions that we don't actually take would have were we to take them. Causality at the most fundamental level concerns action, and I have not yet provided a means to represent action in the formalism. I will now introduce such a means, but first a note about representing observation.

Seeing: The Representation of Observation

The structure of a causal graph, in combination with the probability distribution across its nodes, determines what inferences we can make on the basis of new information. When this information takes the form of an observation, when we see or hear or learn about in some other way the value of a variable in our model, then we must change the probabilities of all other variables, given this new information. This process of changing probabilities to reflect new

information is called updating, and a simple theorem of probability theory can be used to do it.

The theorem was first discussed by the Reverend Thomas Bayes in 1763 in a letter entitled ''An Essay Towards Solving a Problem in the Doctrine of Chances''; therefore, it is generally referred to as Bayes' rule. One school of thought among philosophers is that probabilities are essentially representations of belief, that probabilities reflect nothing more than people's degree of certainty in a statement. Even an assertion that would seem to have a strong objective basis, like the belief that a randomly chosen die is fair—that the probability of each of its six faces coming up when rolled are the same and equal to exactly 1/6—can't be fully justified by analysis or by experience. It can't be proven by analyzing the die because, after all, the faces cannot all be *exactly* the same size and the distribution of mass within the die will ever so slightly tilt the die in favor of one face or another. You could roll the die many times in order to see empirically if the faces have equal probability. But to answer the question with any confidence would require a huge number of rolls, so many that the die would have changed shape by the time you had collected enough observations. The change in shape may be very slight, but even a tiny change could affect the probability of the die coming up 1 or 6 or something else.

You could argue that you don't know which face is more or less likely, and therefore they should all be given equal probability. I think this is a good argument, but it's a claim about what you don't know, not about the way the world really is, and in that sense it's saying that probabilities are grounded in knowledge, in a subjective mental state. For this reason, philosophers who believe that probabilities are representations of belief are sometimes called subjectivists because they claim that probabilities are grounded in subjective states of belief. Sometimes these philosophers are called Bayesian because Bayes' rule turns out to be extremely important on their view.

One argument that has been raised against subjectivism is that if probabilities are grounded in the beliefs of a judge, then it would seem that the judge can make the probabilities anything they want them to be. If a subjectivist wants the probability that a particular horse will win the race to be high, then what is stopping him or her from simply believing and asserting that the probability is high?

The answer is that even a subjectivist's probability judgments have to make sense. They have to *cohere* with all their other judgments of probability, and they have to correspond with what is

known about events in the world. This is how Bayes' rule helps; it prescribes how to take facts in the world into account in order to revise beliefs in light of new evidence. That is, it dictates how to change your probability judgments in the face of new information—how to update your beliefs or, in other words, to learn—while maintaining a coherent set of beliefs, beliefs that obey the laws of probability.

Other schools of thought about the foundations of probability also exist. The most prominent alternative to subjectivism is frequentism, which supposes, roughly speaking, that a probability reflects a long-run relative frequency of an event (the probability of a die landing on 6 reflects the proportion of times it would land on 6 if it were tossed an infinite number of times). But it's beyond the scope of this book to discuss the foundation of probability in any detail.[1] I'll just point out that Bayes' rule isn't specific to subjectivism. It's a rule about how probabilities relate that holds whatever your philosophy of probability.

Bayes' rule looks like this. Say you have some hypothesis about the world, perhaps that your friend Tatiana has a bacterial infection. Call it D (for *d*isease) and its antithesis, that Tatiana does not have a bacterial infection, ∼D (∼ can be read as *not*). Assume you have some degree-of-belief that D is true, P(D) [P(∼D) is just 1−P(D)]. P(D) is often called a prior probability of D because it represents your degree-of-belief before you come across new evidence about Tatiana. Say you do encounter new evidence relevant to D; for instance, it turns out that Tatiana has a peptic ulcer. Call this new datum S (for *s*ymptom). You also have some belief about the effects of D: peptic ulcers are a symptom of bacterial infections. But you know that peptic ulcers can arise for other reasons, ∼D. For example, they might also arise as a result of consuming too much aspirin.

Now you need to revise your belief in D in light of S. But how should you go about it? You want to know P(D|S): the probability of the disease given the symptom or, in this case, the probability that Tatiana has a bacterial infection, given that she has a peptic ulcer. Bayes' rule tells us what it is:

$$P(D|S) = \frac{P(S|D)P(D)}{P(S)}.$$

P(D|S) is called a posterior probability because it is the probability of the disease *after* taking into account the evidence provided by the symptom. Bayes' rule is very easy to prove using the mathematics

of probability. But conceptually it's not that simple. Why should our belief in D given S be related to the probability of S given D and the prior probabilities of D and S in just this way? Instead of trying to answer this question directly, we'll look at it in a way that makes more intuitive sense.

Let's think about what we want to know (how likely it is that Tatiana has a bacterial infection) in a form that doesn't require that we calculate P(S) directly. Instead of calculating the probability of the hypothesis given the data, we can figure out another quantity: P(D|S) divided by P(∼D|S), the odds of the hypothesis after taking the data into account, called the *posterior odds*. This tells us how much more or less likely the hypothesis is than its complement, P(∼D|S). Knowing the odds that our hypothesis is true rather than false is as useful to us as knowing the probability that our hypothesis is true.

I showed Bayes' rule for P(D|S). But I could have just as easily written it, in a completely parallel form, for the complementary hypothesis that Tatiana does *not* have a bacterial infection:

$$P(\sim D|S) = \frac{P(S|\sim D)P(\sim D)}{P(S)}$$

To calculate the odds, P(D|S)/P(∼D|S), all we have to do is divide the first equation by the second (this is the only bit of algebra I'll do in this book!). This gets rid of P(S) by dividing it out:

$$\frac{P(D|S)}{P(\sim D|S)} = \frac{P(S|D)\quad P(D)}{P(S|\sim D)P(\sim D)}$$

That's enough mathematics. You can think about this equation as having three parts: the posterior odds, P(D|S)/P(∼D|S); what I'll call the *likelihood ratio,* P(S|D)/P(S|∼D); and the *prior odds*, P(D)/P(∼D). The prior odds is just the ratio of our prior probabilities, the odds that the hypothesis is true before taking the evidence into account. The equation shows that the three parts are related in a very reasonable way:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}.$$

Our belief after incorporating the new evidence S (the posterior odds) is equal to whatever our belief was before (the prior odds) times the likelihood ratio.

The likelihood ratio tells us how strong the evidence S is for or against D. Specifically, it tells us how well S distinguishes D from

~D. Tatiana has some probability of having a peptic ulcer by virtue of D, a bacterial infection. But she also has some probability of having a peptic ulcer by virtue of ~D, having consumed too much aspirin or for some other reason. The higher the former probability and the lower the second, the higher the likelihood ratio will be. The likelihood ratio is the odds that the new fact (Tatiana has a peptic ulcer) would arise if the belief we're concerned about (Tatiana has a bacterial infection) is true relative to it being false. Learning about the peptic ulcer will help us diagnose whether Tatiana has a bacterial infection only if peptic ulcers are either much more probable given bacterial infections than given other reasons, or if they are much less probable. If their probabilities are roughly the same, then the likelihood ratio will be close to 1 and learning about the peptic ulcer won't tell us much.

The likelihood ratio requires knowing P(S|D) and P(S|~D). These are quantities we can usually obtain. If we have some belief about the way the world is (knowledge about what produces peptic ulcers), then we generally know what to expect when various diseases occur (the probability that a bacterial infection will result in a peptic ulcer and that other conditions will lead to a peptic ulcer). The likelihood ratio is really important and valuable because it is so simple yet it tells us so much. It tells us how well the data—new facts—discriminate one belief from another, and when you're deciding what to believe from observations, that's what determines the value of information.

Let's summarize how we learn about Tatiana from observation. We start with some belief that Tatiana has a bacterial infection, denoted by the prior odds. We then learn that she has a peptic ulcer. So we multiply the prior odds by the ratio of the probability that she has a peptic given that she has a bacterial infection with the probability that she has a peptic ulcer for some other reason (because she's consumed too much aspirin, say). If this ratio is greater than one (we believe that peptic ulcers are more often caused by bacterial infections than aspirin), we end up with a posterior odds greater than the prior; we increase our belief that she has a bacterial infection. But if the ratio is less than one because we think peptic ulcers are more likely to be a result of aspirin, we lower our posterior odds; we decrease our belief in bacterial infection.

This logic based on Bayes' rule is natural and may be familiar to you. We start with a degree-of-belief in the way the world is; we look at the world; if the world is consistent with our beliefs, then we increase our degree-of-belief. If it's inconsistent, we decrease it.

In fact, if we do this long enough, if we have enough opportunity to let the world push our probabilities around, then we would all end up with the same probabilities by updating our beliefs this way. Bayes' rule is guaranteed to lead to the same beliefs in the long run, no matter what prior beliefs you start with.

Notice how important symptoms are for determining disease according to this logic. Our beliefs tell us which diseases lead to which symptoms. But by learning about someone's symptoms (observables), we can make good guesses about the disease they have (an unobservable). This is the technical sense of *diagnostic*: effects (symptoms, in this case) help us figure out the causes (diseases, in this case) that must have been present through the kind of backwards inference exemplified by Bayes' rule.[2]

Action: The Representation of Intervention

Presumably the point of all this belief formation and updating is to know how the world works, what the causal mechanisms are that guide it. In chapter 3, we learned that the way to find out about causal structure is not through mere observation but through experiment and what distinguishes the two is that experimentation requires action; it requires intervening somewhere in the system. Imagine that instead of observing values, we bring them about ourselves through our action. What if we get rid of Tatiana's peptic ulcer by giving her Grandma's special formula? All we know about Grandma's special formula is that it goes directly to the ulcer, bypassing all normal causal pathways, and heals it every time (unfortunately, Grandma passed away taking her formula with her or we'd be rich). Before our intervention, our causal model of peptic ulcers looked something like this:

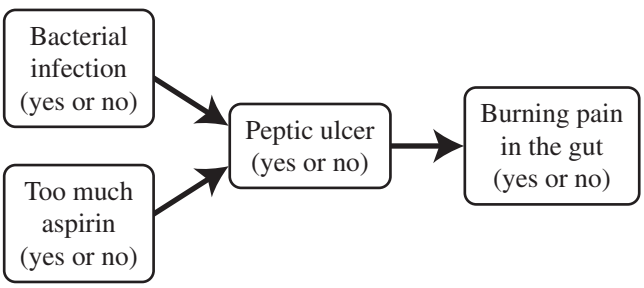

Figure 5.1

Our direct intervention with Grandma's special formula warrants a change in our belief that Tatiana has a peptic ulcer (i.e., we should now believe that she doesn't), but because we're the ones that got rid of her peptic ulcer, we had no effect on the normal causes of that ulcer. Tatiana is just as likely to have a bacterial infection or to have consumed too much aspirin as she was before we applied Grandma's formula. Therefore, we should not change our belief in the probability of those causes. We should not make a diagnostic inference from the absence of Tatiana's peptic ulcer to the presence of bacteria or her consumption of aspirin because those *have nothing to do with the absence of the peptic ulcer*. We should act as if the normal causes of peptic ulcer are independent of the ulcer because they no longer have an effect; we're the ones that got rid of the ulcer.

Graphically, the relevant causal model after intervention temporarily looks like this:

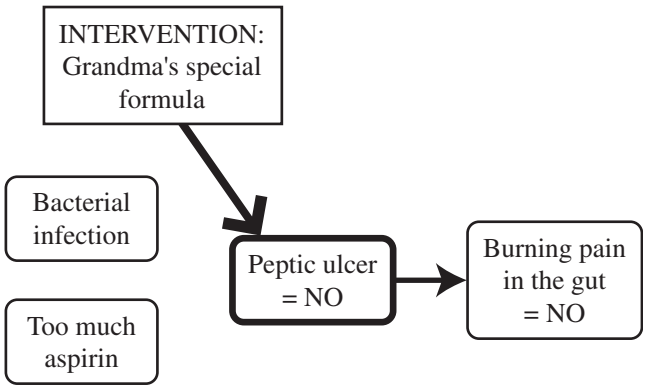

Figure 5.2

Our intervention on peptic ulcer had three effects. First, it set the value of peptic ulcer to NO because Tatiana no longer has one. Second, and less obviously, it changed our causal model by cutting peptic ulcer off from its normal causes so that the absence of the ulcer would not be treated as diagnostic of those causes. Third, because the intervention did not cut peptic ulcer off from its effects, only its causes, it also reduced Tatiana's burning pain in the gut. After all, if Grandma's magic formula gets rid of Tatiana's ulcer, it should also relieve the pain from the ulcer, whatever the cause of her ulcer. We can make this inference because the link from peptic ulcer to burning pain remains intact.

The general lesson is that the inferences that we can draw after *observing* a particular value of a variable are not the same as those that we are licensed to draw after *intervening* to set the variable to that same value (''no ulcer'' in the example). Bayesian updating by itself fails to recognize this difference. We cannot use Bayes' rule to figure out the probability of a cause after intervening to set an effect because the links between cause and effect no longer exist; the intervened-on effect is not diagnostic of its causes.

The absence of peptic ulcer suggests nothing about the presence of a bacterial infection or about aspirin intake after intervention on the peptic ulcer. Of course, we don't need to use Bayes' rule because things are simpler with intervention. The probability of the cause is whatever it was before our intervention; no updating is necessary.

Acting and Thinking by Doing: *Graphical Surgery*

What's new and different about the causal modeling framework is that it gives us a way to represent intervention and distinguish it from observation. To do so, Pearl[3] introduced the *do* operator. To represent intervention on a variable X in a causal model by setting X to some value x, we $do(X = x)$. Previously, we did the equivalent of

$$do(\text{peptic ulcer} = \text{NO}).$$

The *do* operation doesn't only set a variable to a value; it also modifies the causal graph by disconnecting X from its normal causes. Because an agent is acting to determine the value of X, the normal causes of X have no influence on it. The action overrides the causes of X, rendering them irrelevant. Otherwise, the graph remains the same. Pearl calls this a surgery. The graph is surgically altered by removing one specific connection and leaving others intact. The effects of the intervention are then computable through normal probability calculations on this ''mutilated'' graph.[4] In other words, the *do* operator is exactly what we need to take us from the first causal model to the second. As a result, it allows us to represent actual physical intervention (like applying Grandma's special formula). It also allows us to represent intervention in our minds, counterfactual intervention, by imagining what would happen *if* we did something (like apply Grandma's formula, which, if truth be told, doesn't actually exist). It gives us a way to represent another possible world that we might imagine or fantasize about or pretend to live in or use to make an argument (mothers love possible-worlds

arguments: ''If you were more considerate, then people would like you more''. So do professors: ''If you had done the analysis the way I did, then you wouldn't be stuck in this quagmire''). It even tells us how to represent that other possible world. Start with the causal model of the world you're in, choose the aspect of that world that you want to be different, *do* it by changing that aspect as specified by the counterfactual assumption (i.e., set its value to the value imagined), and then cut it off from its causes. You'll end up with a new causal model of the new world you're thinking about, but one that's very similar to the old world. In fact, as long as the variable we're intervening on is not a root node in the causal graph (a variable with no known causes), the new world will be a simplified version of the old world, one with fewer causal links. If it is a root node, we'll end up in a world with an identical causal structure.

The representation of an imagined (counterfactual) intervention is obtained in exactly the same way that the representation of an actual physical intervention is obtained. Hence, the two graphs of Tatiana's medical concerns before and after intervention with Grandma's special formula illustrate the inference process, even if we're only imagining what would happen if it were applied. The only real difference in reasoning about actual versus counterfactual intervention is that the graph after intervention represents the actual world in the case of actual intervention but another possible world in the case of counterfactual intervention. To reason counterfactually, you make an assumption. You might imagine that Tatiana has no peptic ulcer, *do*(peptic ulcer = NO); or that the moon were made of cheese, *do*(moon's composition = CHEESE); or that Tatiana loved me, *do*(Tatiana's love interest = STEVEN); and so on. Then trace through the implications of the assumption to see what its effects are (the moon would be covered with mice and people who love cheese, etc.). What you don't do is change the normal *causes* of the facet of the world whose value you're assuming, only its effects. You don't assume that the big bang led to a lot of cheese pervading the universe that coalesced into the moon. Such an assumption would be irrelevant. Instead, you consider the consequences of a cheesy moon.

Science fiction authors do this all the time. They make an assumption, often known to be false (there is life on Venus; everyone has access to a mind-altering wonder drug without any side effects), and draw out the implications of their assumption. To question the validity of the assumption would involve pointing out how the normal causes of that aspect of the world would never lead to the assumed state (''Venus is too hot to support life and it has no

water anyway''), but that would be uncooperative and seems pedantic. We assert things all the time in order to capture their effects (''If only there was life on Venus, then I could find a friend''). Questioning whether the assumption is valid fails to address the issue. What matters is what that asserted world would be like, not whether it is possible.

In the simplest cases, the disconnection involved with the *do* operator separates a small chunk of a causal model from the rest of a larger model so that we can limit our stream of inferences to the smaller subset. The *do* operator has the effect of biting off a small part of our big model of the world and limiting our thinking to that part. Thus, we can use the *do* operator to think about a fictional world without changing our beliefs about the real world. Or we can use it to help us focus on some relatively small problem, like how to fix our car or what the results of an experiment should be, without worrying for the moment about the much bigger and more complicated world outside our current problem.

Computing With the Do *Operator*

Once we have the *do* operator, we can ask questions like what would be the probability of a bacterial infection if I eliminated the ulcer using Grandma's special formula? Using the *do* operator, this would be represented as

$$P(\text{Bacterial infection} \mid do(\text{peptic ulcer} = \text{NO})),$$

or what would be the probability of burning pain in the gut if I eliminated the peptic ulcer,

$$P(\text{burning pain in the gut} \mid do(\text{peptic ulcer} = \text{NO}))?$$

These are questions about *interventional* probabilities rather than *conditional* probabilities. The first interventional probability is equal to the prior probability of bacterial infection,

$$P(\text{Bacterial infection} \mid do(\text{peptic ulcer} = \text{NO})) = P(\text{Bacterial infection})$$

because peptic ulcer has been disconnected from bacterial infection and so provides no information about it. The second is identical to the corresponding conditional probability because the intervention has no effect on the relevant causal link:

$$\begin{aligned} &P(\text{burning pain in the gut} \mid do(\text{peptic ulcer} = \text{NO})) \\ &\quad = P(\text{burning pain in the gut} \mid \text{peptic ulcer} = \text{NO}). \end{aligned}$$

How do we calculate interventional probabilities in general? As we've just seen, we can read them off a fully specified causal model. In the example, if we know P(Bacterial infection) and we know the causal model as shown, then we know P(Bacterial infection| *do*(peptic ulcer = NO)). The trick is to reduce interventional probabilities to some combination of probabilities and conditional probabilities. We have to be careful if there are what Pearl calls "backdoor paths." This occurs when some other causal route links bacterial infection to peptic ulcer even after the links to the causes of peptic ulcer are removed, as would be the case if this were the operative causal model:

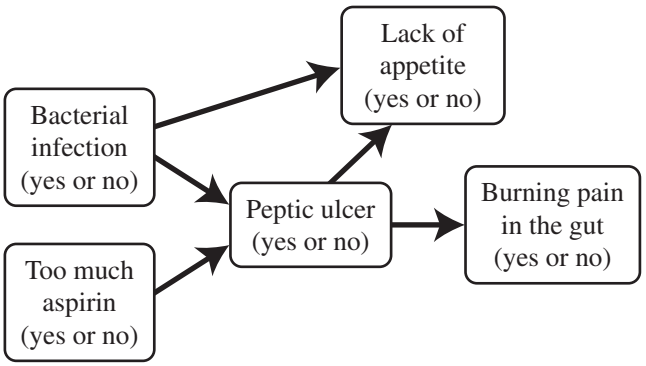

Figure 5.3

According to this model, peptic ulcers and bacterial infections are both causes of lack of appetite. So if you intervene to get rid of someone's peptic ulcer, you're likely to change their appetite. But a change in appetite can be informative about its causes, one of which is bacterial infection. So because of the backdoor path between bacterial infection and peptic ulcer that goes through appetite, intervening on peptic ulcer does not render it independent of bacterial infection. In such a case, extra precautions need to be taken to evaluate interventional probabilities in terms of conditional probabilities using the model. Calculations of interventional probabilities can get difficult, although there are software packages that can help.

Another way to evaluate an interventional probability is to run an experiment. If we collect a group of people and give them Grandma's special formula, then as long as Grandma's formula does not produce bacterial infection as a side effect, we can find out how many have bacterial infections, and that's our answer regardless of any backdoor paths.

The Value of Experiments: A Reprise

Now that I've spelled out how to represent observation and how to represent intervention, I can say more about observational studies, scientific studies that tell us only about whether variables are correlated or not and compare them more fully to experiments, which involve intervention. First, we see that the primary importance of observational studies is that they tell us whether variables are dependent or independent. And if we have enough data that we give sufficient credibility to, we can even tell if variables are conditionally independent or not. So there's a lot of information in correlational studies that measure the right variables. They can really narrow down the correct causal model. For any given three variables (e.g., intelligence, socioeconomic status, and amount of beer consumed per week), we can reduce the number of possible relations among them to three or four. For example, as we saw in the last chapter, if two of the variables are dependent, say, intelligence and socioeconomic status, but conditionally independent given the third variable, then either they are related by one of two chains

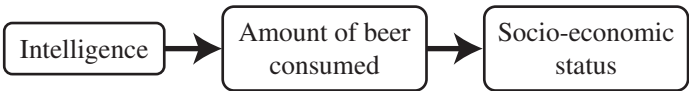

Figure 5.4

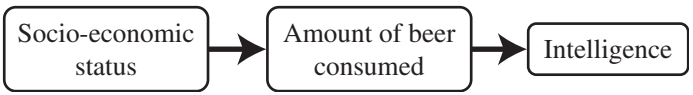

Figure 5.5

or by a fork

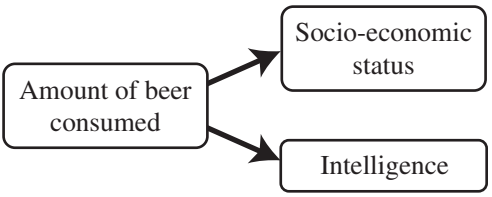

Figure 5.6

and we then must use other means to decide between these three possibilities. In some cases, common sense may be sufficient, but

we can also, if necessary, run an experiment. If we intervene and vary the amount of beer consumed and see that we affect intelligence, that implies that the second or third model is possible; the first one is not. Of course, all of this assumes that there aren't other variables mediating between the ones shown that provide alternative explanations of the dependencies.

If we choose our variables wisely, then every correlation derives from a causal model of some sort. In general, a dependency between two random variables A and B could result from any of the following causal models.

Figure 5.7

Figure 5.8

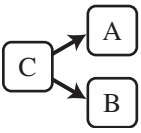

Figure 5.9

So we need a fair number of correlations, correlations conditional on other variables, and experiments to allow us to infer the right causal model given the large number of variables in most real causal systems. That's why learning causal systems can be hard. Indeed, some causal systems, like the solar system, required thousands of years of intense study before they were figured out, in part because nobody was able to run an experiment to vary the movement of the planets! But even when we can run experiments, causal systems can elude description. How the mind works, for example, is still a wide open question. Think of all the relevant variables. At the lowest level of description, everything that a mind can ponder is a cause of the mind's operation, and every thought and action that the mind can produce is an effect.

The Causal Modeling Framework and Levels of Causality

The conception of intervention that I've developed in this chapter draws on the idea that the directed links in a causal graph represent

local, autonomous, and stable mechanisms. Intervening to set a variable to a fixed value disrupts those mechanisms that previously controlled that variable but leaves all other mechanisms intact. Further, it formalizes the intuitive notion that X causes Y (directly or indirectly) if some subset of interventions to change X *would* lead to a change in Y. It is this synergy of the concepts of causal link, mechanism, and intervention that gives the causal model framework its originality and power.

I have been talking as if the world is composed of one very large, very complicated set of causal mechanisms. But in fact the world is composed of many such sets because we can talk about causal systems in the world at different hierarchical levels.

A causal graph represents a causal system at a particular level of granularity. Causal systems afford multiple descriptions, though, at multiple levels of description. The human body can be described coarsely, in terms of its physiological systems (nervous system, cardiovascular, etc.) and how they interrelate, or, more finely, each physiological system could be decomposed into its parts. Clearly, this could be done at multiple levels of precision. The most specific level describes a particular causal event in terms of all of its contingent subevents, all relevant variables and their interactions and effects. One view is that a causal analysis at this level is deterministic because it describes every causal factor without ignoring anything (for the sake of discussion, we leave aside quantum indeterminism, as it probably has little to do with everyday causal models). Of course, descriptions this precise may not be possible, so such deterministic models may be merely ideals. As soon as variables are ignored, descriptions become more coarse. We almost always do ignore variables. If we ask whether smoking causes cancer, and certainly if we do so in a nonscientific context, we tend to ignore other contaminants in the environment, what people are eating, the shape of a person's lungs, all factors that could well contribute to cancer and to smoking's effect on cancer. This is when causality becomes probabilistic.

Causal analyses are hierarchical in another sense as well, in terms of their level of abstraction. A particular causal relation at its most specific level of description is a realized mechanism; at its most general it is a causal principle for generating mechanisms. For example, a particular guillotine is a specific causal mechanism. The mechanism depends on some abstract causal principles, one being that gravity causes acceleration. Note that general principles apply at every level of granularity, and, in this sense, these two kinds of

hierarchy are independent. A plausible psychological hypothesis is that people store relatively few causal models, and certainly few at a detailed level of analysis. Instead, people may store general causal principles that allow them to construct causal models—and thus explanations for events—on the fly. As a result, people's causal models may usually be quite impoverished. Psychologist Frank Keil argues not only that they are but that people don't know just how impoverished they are![5]