2

Spurious correlation and probability increase

The first main qualification of the basic probability-increase idea of probabilistic causation, explained in Chapter 1, is the relativity of the causal relation to a given *token population,* considered to be of a given (appropriate) *kind* that the population exemplifies.[1] The second main qualification of the basic probability-increase idea, to be explored in this chapter, involves the possibility of what has been called "spurious correlation." Of course, what is meant by saying that a factor X raises the probability of a factor Y is that $Pr(Y/X) > Pr(Y)$ – equivalently, $Pr(Y/X) > Pr(Y/\sim X)$.[2] Another way of expressing this relation is to say that Y is positively probabilistically *correlated* with X. It is famous that "correlation is no proof of causation," and it is also true that causation does imply correlation. The possibility of spurious correlation is one reason why.

In this book, I will actually explore in detail three general

[1] As noted in Chapter 1 (note 18), the actual token population washes out, or disappears, so to speak, when probability is analyzed in terms of hypothetical relative frequencies involving infinitely many populations or individuals. The actual frequencies in a finite token population have no mathematical effect on a hypothetical infinite limit. All that is relevant, on this kind of interpretation of probability, is the *kind* that the actual token population is considered to exemplify. But, of course, there are other ways of understanding probability than hypothetical limiting-frequency approaches. In any case, in what follows, I will often include (perhaps needlessly) the actual *token population* (as well as its *kind*) as a relatum in the relations of probability and of population level probabilistic causation.

[2] $Pr(Y) = Pr(X)Pr(Y/X) + Pr(\sim X)Pr(Y/\sim X)$. So, assuming that $Pr(X) \neq 0$ and $Pr(\sim X) \neq 0$, $Pr(Y)$ is an average of $Pr(Y/X)$ and $Pr(Y/\sim X)$, and must therefore lie strictly between these two values. Henceforth, I will for the most part use the letters "X," "Y," "F," "G," and so on to refer to factors, rather than "C" and "E," which already suggest "Cause" and "Effect," which can perhaps be misleading in some cases.

ways in which probability increase may fail to coincide with causation, and I will show how the probability-increase idea of causation should be adjusted to accommodate these three possibilities. After briefly describing the three possibilities below, this chapter will concentrate on one of them, the one called "spurious correlation." The other two will be dealt with in subsequent chapters.

One simple way to see that probability increase does not imply causation is to notice that the relation of positive correlation is symmetric. If X raises the probability of Y, then Y raises the probability of X.[3] So if probability increase implied causation, then causation would be symmetric as well, at least for probability-increasing causes. But clearly the relation of causation is not symmetric, and we do not want our theory to imply that Y is a cause of X whenever X is identified as a cause of Y. In fact, most plausibly, the relation of causation is *asymmetric.* If so, then if X causes Y, then Y *does not* cause X. So if X is a probability-increasing cause of Y, then Y is a *probability-increasing noncause* of X.

A natural approach to this kind of probability increase without causation would be to include in the theory of probabilistic causation, along with the probability increase idea, a condition requiring that a cause precede its effect in time. This would handle this kind of probability increase without causation, because temporal precedence is not symmetric. In Chapter 5, I argue that the temporal priority idea *must* be *explicitly* incorporated into the probabilistic theory, in order to handle what we may call this "problem of temporal priority of causes to effects." Until Chapter 5, let us adopt the convention that the factor denoted by the letter "X" is temporally prior to the factor denoted by the letter "Y." In most cases this will be obvious. But the idea of one *factor* (or *property* or *type,* an *abstract* thing) preceding another in time is somewhat

[3] If $Pr(Y/X) > Pr(Y)$, then, by the standard definition of conditional probability, $Pr(Y\&X)/Pr(X) > Pr(Y)$. It follows that $Pr(Y\&X)/Pr(Y) > Pr(X)$, so that again by the definition of conditional probability, $Pr(X/Y) > Pr(X)$.

subtle and puzzling. In Chapter 5, I suggest a way of understanding this idea that is very natural in light of the relativity of probability and probabilistic causation to populations.

A second kind of probability increase without causation involves the possibility of what has been called "causal interaction." This kind of possibility, and its relevance to the probability-increase theory of type-level causation, is easy enough to explain. But in order to motivate the solution to the problem adequately, it is necessary first to lay down some of the rudiments of the theory, the relevant parts of which will be given in this chapter. The following may provide a general idea of the problem, which is addressed in detail in Chapter 3.

It is possible for a causal factor X to *interact* with a factor F, relative to the production of a factor Y, in the sense that the causal significance of X for Y is different when F is present from what it is when F is absent. To use an example of Cartwright's (1979), ingesting an acid poison (X) is causally positive for death (Y) when no alkali poison has been ingested ($\sim F$), but when an alkali poison has been ingested (F), the ingestion of an acid poison is causally negative for death. I will argue that in a case like this it is best to deny that X is a positive causal factor for Y, even if, overall (for the population as a whole), the probability of death when an acid poison has been ingested is greater than the probability of death when no acid poison has been ingested (that is, even if $Pr(Y/X) > Pr(Y/\sim X)$). I will argue that it is best in this case to say that X is causally *mixed* for Y, and despite the *overall* or *average* probability increase, X is nevertheless not a positive causal factor for Y in the population as a whole.

Chapter 3 shows that another problem that arises in thinking about causation, the problem of disjunctive causal factors, is an instance of this kind of problem. First, however, we must deal with a third kind of probability increase without causation, which has been called "spurious correlation." The resolution of this problem will give us the framework, and

part of the motivation, for the resolution of the problem of probabilistic causal interaction.

2.1 SPURIOUS CORRELATION

On what seems to be the usual understanding of the term, two factors are *spuriously correlated* when (roughly) neither causes the other and the correlation disappears when a third variable is introduced and "held fixed" – that is, the correlation disappears both in the presence and in the absence of the third factor.[4] This does not quite capture the kind of situation I will explore in this chapter. As Simon (1954) is careful to point out, if a correlation between factors X and Y disappears both in the presence and in the absence of a third factor Z, then the explanation may be *either* that the correlation results from the joint causal effect of Z on X and Y (Z is a common cause of X and Y) *or* that Z is an intermediate causal factor between X and Y (X operates on Y through Z or Y operates on X through Z). We shall not count the second possibility as a case of spurious correlation. In the second case, one of X and Y may in fact be a genuine positive causal factor for the other (of course, given our convention that the factor represented by the letter "X" temporarily precedes the factor represented by the letter "Y," it cannot be that Y causes X). This kind of case will be discussed in Chapter 4, on causal intermediaries and transitivity of causal chains.

So let us for now understand there to be a spurious correlation between two factors X and Y if neither causes the other and they are correlated effects of a common cause Z, where the correlation of (the later) Y with (the earlier) X disappears when Z is held fixed. Because we are excluding the case in which Z is causally intermediate between X and Y, I sometimes refer to the common cause Z as a "separate cause" of factor Y – that is, a cause of Y that is separate from X's causal

[4] See, for example, Simon (1954), Suppes (1970), and Skyrms (1980).

role, if any, for Y. This understanding of spurious correlation will have to be generalized in several ways below, but first some explanation of the definition and some examples to illustrate the idea.

Suppose a factor Z is a cause of both X and Y. See Figure 2.1. (In this figure, and in others that follow, the solid lines with arrows represent causal connections, the broken lines represent correlations, and the "+"'s and "−"'s indicate whether the causal impact or the correlation is positive or negative.) In the simplest kind of common cause case (others will be considered later), the following relations hold:

(1) $Pr(X/Z) > Pr(X/\sim Z)$,
(2) $Pr(Y/Z) > Pr(Y/\sim Z)$,
(3) $Pr(Y/Z\&X) = Pr(Y/Z\&\sim X)$,
(4) $Pr(Y/\sim Z\&X) = Pr(Y/\sim Z\&\sim X)$.

Propositions (1)–(4) imply $Pr(Y/X) > Pr(Y/\sim X)$.[5] (1) and (2) correspond to the assumption that Z is a common cause of X and Y, on the probability increase idea. (3) and (4) say what it means for the correlation between X and Y to disappear when Z is held fixed (positively and negatively); and they correspond roughly to the assumption that Z is the *only* factor involved that has any causal influence on any of the others. And the derivation of $Pr(Y/X) > Pr(Y/\sim X)$ from (1)–(4) is supposed to *explain* (in the simple kinds of cases I have in mind now) the correlation of Y with X in terms of the "screening off" common cause Z.[6]

[5] $Pr(Y/X) = Pr(Z/X)Pr(Y/Z\&X) + Pr(\sim Z/X)Pr(Y/\sim Z\&X)$. So, by (3) and (4),
$Pr(Y/X) = Pr(Z/X)Pr(Y/Z) + Pr(\sim Z/X)Pr(Y/\sim Z)$.
Also by (3) and (4),
$Pr(Y/\sim X) = Pr(Z/\sim X)Pr(Y/Z) + Pr(\sim Z/\sim X)Pr(Y/\sim Z)$.
Let $a = Pr(Z/\sim X)$ and $b = Pr(Y/\sim Z)$. Then by (1) and (2), and symmetry of correlation, there are positive numbers u and v such that $Pr(Y/X) = (a + u)(b + v) + (1 - a - u)b = av + uv + b$, and $Pr(Y/\sim X) = a(b + v) + (1 - a)b = av + b$. Since $uv > 0$, $Pr(Y/X) > Pr(Y/\sim X)$.

[6] This is the idea articulated by Reichenbach (1956) in his "principle of the common cause," according to which correlated factors can be explained in terms of a common cause. (3) and (4) are what it means to say that Z and $\sim Z$ each *screen off* Y from X – the same as what it means to say that the correlation between X and Y disappears when Z is held fixed. Salmon (1978) calls the kind of structure described by (1)–(4) a "conjunctive fork," which he distinguishes from "interactive forks," in which one or both of (3) and (4) are false. I discuss interactive forks in Section 2.4. For criticisms of the principle of the common cause, and assessments of the purposes to which it has been put, see also van Fraassen (1977b, 1980), Salmon (1984), Sober (1984b) and (1987a), Torretti (1987), and Forster (1988).

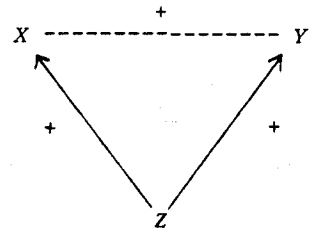


Figure 2.1

Here is a concrete example that is often used to illustrate this: The British statistician Ronald Fisher (1959) once considered the possibility that lung cancer (Y) is positively correlated with smoking (X) not because smoking causes cancer, but because there is a genetic common cause (Z) of the two. In this hypothetical example, X increases the probability of Y, even though X in no way causes Y. The probability increase is due to smoking's increasing the probability of its cause, the "smoking gene," which in turn increases the probability of the gene's effect, lung cancer. But if we hold fixed whether or not the gene is present, smoking will not increase the probability of lung cancer. A different example of this was discussed in Chapter 1. Rain (Y) is correlated with falling barometers (X) that precede the rain. But falling barometers do not cause rain. Again, the two factors are effects of a common cause: approaching cold fronts (Z).

As I mentioned above, our understanding of spurious correlation must be generalized; the characterization given above is too narrow. For one thing, there is the possibility of a spurious correlation arising from the operation of *multiple* separate causes of a factor Y, where these causes are causally indepen-

dent of each other and of the factor X whose causal role for Y is in question. The analysis of this kind of case will be somewhat different from that of the single separate cause (Z) case. I will turn to the multiple separate causes case at the end of this section, and will for now restrict the analysis to the case of the single separate cause.

The more general phenomenon of spurious correlation that I characterize below can arise in a number of ways, with various possibilities for the true causal relation between X and Y, including but not limited to causal neutrality. The basic idea behind the more general understanding of spurious correlation can be expressed, intuitively, like this: Because of the operation of a factor commonly causally relevant to X and Y, the *magnitude* of correlation of Y with X is different from the magnitude of X's causal significance for Y. The idea is to include as cases of spurious correlation, not only cases in which the *direction* of inequality between $Pr(Y/X)$ and $Pr(Y/\sim X)$ fails to coincide in the natural way with the kind of causal significance X has for Y, but also cases in which the magnitude of the difference between $Pr(Y/X)$ and $Pr(Y/\sim X)$ fails to coincide with the degree of causal significance of X for Y. Before being more precise about this (in particular, about how we should understand these magnitudes), I will illustrate the idea with the help of a well-known example, and variations on it.

Nancy Cartwright (1979) cites a study by Bickel, Hammel, and O'Connell (1977) on graduate admissions at Berkeley. It was found that, in the population of all applicants, getting admitted was positively correlated with being male. The frequency of admission among male applicants was higher than the frequency of admission among female applicants. This naturally suggested discrimination against women, and (as Cartwright puts it) "thus rais[es] the question '*Does being a woman cause one to be rejected at Berkeley?*' " (Equivalently: "Is being male a positive causal factor for getting admitted at Berkeley?") However, admissions decisions were made

within the academic departments to which one applied. And when the admissions histories of the departments were investigated separately, one by one, it was found that there was no department within which there was a correlation between gender and getting admitted. This is consistent with the fact that, *on average,* the frequency of admission was lower among women than it was among men. The women applicants tended to apply to departments into which it was harder to gain admission.

The table below gives an example of how this can happen; all the entries are the number of accepted applicants over the number applying.

	Department 1	Department 2	Total
Male	81/90	2/10	83/100
Female	9/10	18/90	27/100

In this example, Department 1 accepts 90 percent of all male applicants as well as 90 percent of all female applicants, while Department 2 accepts 20 percent of all male applicants as well as 20 percent of all female applicants. Within each department, there is no correlation between gender and admission. Overall, however, the probability of getting admitted is more than three times greater for male applicants than it is for female applicants. The department-by-department analysis of admissions records was taken as exonerating the Berkeley graduate school from the charge of discrimination: Being male, it now seems, was not after all a positive causal factor for getting admitted, in the population of all applicants to the Berkeley graduate school.

If the question of discrimination against women in the example is equivalent to the question "Is being male a positive causal factor for getting admitted?," and if the more careful look at the data in fact shows that Berkeley is not guilty of

discrimination, then we have to conclude that being male is not, after all, a cause of getting admitted. However, there is presumably *no common cause* of being a male and getting admitted. So this seems to be a case of a factor Y (getting admitted) being positively correlated with a factor X (being a male) where X does not cause Y (and of course Y does not cause X), *yet there is no common cause Z of X and Y*. Is this a new kind of correlation without causation, not of any of the kinds described above?

If we look more carefully at the example, it turns out that it really is of the common cause kind. First, let us ask what is responsible for the correlation between *being male* and *applying to a department that is relatively easy to get into*. Of course, it would be implausible to suppose that the latter causes the former or that there is a common cause of the two. Most plausibly, being male somehow causes one to apply to the departments that are relatively easy to get into (possibly because of the way males tend to be brought up, getting them interested in the subjects taught in the larger, better funded departments, for example).

Second, I want to question the assumed equivalence between *there being discrimination against women* and *being male's being a positive causal factor for admission*. Although this is perhaps a fine point, it actually does make a crucial difference in the analysis of this example. If an institution *is* guilty of discrimination against women, then it is not, strictly speaking, *being* male that is necessarily a positive causal factor for admission, but rather *the institution's believing of an individual that he (or she) is male* that is the positive causal factor.

Suppose an institution in fact *does* have a policy of discriminating against women. Suppose also, as a thought experiment, that one year all the men were persuaded to check "female" on their applications, and all the women "male." Then we would expect that *being* male could be a negative causal factor for admission. Nevertheless, the charge of discrimination against women holds up, because the institu-

tion's *believing* of an individual that he or she is male *is* a positive causal factor for admission. As another thought experiment, suppose that the institution does have a policy of discrimination against women and that one year all the applicants decided in some random way whether to check "male" or "female" on their applications. Then it could be that actual gender is causally neutral for admission, despite the policy of discrimination against women.

In the Berkeley example, I think that being male *is* a positive causal factor for admission and that the graduate school is *not* guilty of discrimination. Being male is causally positive for admission; being male causes one to apply to departments that are relatively lenient in their admissions policies, which in turn is causally positive for admission – and transitivity of causation is plausible in this case. But there is no discrimination: There is no correlation, within any department, between admission and the department's believing of an applicant that he (or she) is male. And it is *this*, strictly speaking, that the department-by-department analysis of admissions history must have turned up, in order for it to be correct to conclude from the study that there is no discrimination against women applicants at Berkeley. Also, strictly speaking, the rows in the table earlier that describes the example should correspond to *being believed to be* male and *being believed to be* female.

Now let X be the factor of the Berkeley graduate school (or a department) believing of an applicant that he or she is male, let Y be the factor of getting admitted, Z the factor of being male, and W the factor of applying to one of the departments that are relatively easy to get into. Then the rows in the table should be relabeled X and $\sim X$, and the causal structure of the example is as diagrammed in Figure 2.2. Of course, Z is causally positive for X. Now if the only way in which Z can affect Y is by way of its influence on W (which seems plausible given the description of the example), then (as will be shown in Section 4.3) probabilistic type causation from Z to

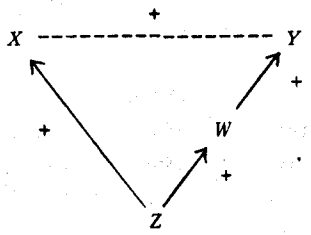


Figure 2.2

W to *Y* will be *transitive: Z* causes *Y*. Thus, Figure 2.2 depicts a special case of the common cause structure shown in Figure 2.1. Here, being male really is a cause of getting admitted, but getting admitted is only spuriously correlated with the school's believing of an applicant that he or she is male: *X* is causally neutral for *Y*.

The idea of *Y*'s being spuriously correlated with *X* should be consistent with the causal significance of *X* for *Y* being other than neutrality. For example, *X* could be *causally negative for Y*, consistent with the more general idea of spurious correlation mentioned above and formulated more precisely below. To see this, consider this slight modification of the example just described. Suppose that in each department there is a lower frequency of admission among applicants believed to be male than there is among those believed to be female: suppose there is a certain amount of "reverse discrimination" in each department, so that *X* is causally negative for *Y*. Still, if the tendency of women to apply to departments that are harder to get into is sufficiently strong, then there will remain a positive correlation between being believed to be male and getting admitted. For an example, simply change the entries in the top row of the table above that describes the original Berkeley example to read, "45/90, 1/10, 46/100." In this example, being believed to be male *increases the probability,* overall, of getting admitted, yet being believed to be male

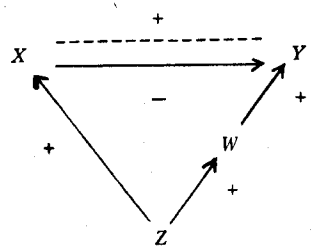


Figure 2.3

is a *negative causal factor* for getting admitted. This situation is diagrammed in Figure 2.3.

Brian Skyrms (1980) has described another example with basically this same feature. Suppose that air pollution in the cities got so bad that city-dwellers tended to refrain from smoking, so as not to put their lungs in double jeopardy. And suppose that people who lived in the country, where there is little pollution, generally felt safe enough to indulge. If the air in the cities is bad enough, and if the ratio of smokers in the cities to smokers in the country is low enough, then the frequency of lung disease could turn out to be lower among the smokers than it is among the nonsmokers. This is because the smokers tend to live in the country, where the air is clean, and the nonsmokers tend to live in the cities, where they are exposed to severe air pollution.

Nevertheless, of course, smoking (as well as exposure to air pollution) is causally positive for lung disease. In this case, a factor *X* (smoking) is a cause of a second factor *Y* (lung disease) even though the former lowers the probability of the latter, on average. This is a case of spurious correlation as characterized above, since pulmonary health ($\sim Y$) is positively correlated with, though not caused by, smoking. Also, if we identify living in the country – call this factor *Z* – as a common cause of pulmonary health and smoking, then we have the common cause structure shown in Figure 2.4.

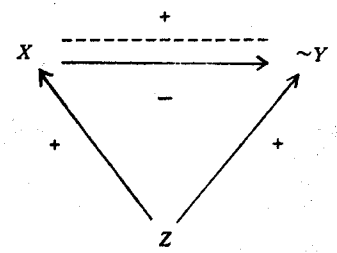


Figure 2.4

Another possibility consistent with spurious correlation is that X is in fact causally *positive* for Y. Here, it is not the *fact* of positive correlation that is spurious (since X *is* a cause of Y), but rather the *magnitude* of the correlation. By the magnitude of Y's correlation with X, I mean simply the difference between $Pr(Y/X)$ and $Pr(Y/\sim X)$. Below, I will be more precise about "spurious magnitudes" of correlation, but first, a couple of examples will illustrate the possibility intuitively, and show further the need to generalize the idea of spurious correlation in a way that takes into account magnitudes of correlation and causal significance.

Consider this modification of the Berkeley admissions example. Suppose there is some discrimination against women in each department, but only very little; within each department there is a small correlation between X and Y. Still there could be a large correlation between X and Y overall. For an example, change the bottom row in the table above used to describe the original Berkeley example to read, "8/10, 16/90, 24/100." In this example, there is a slight tendency of X to cause Y, but a large correlation between X and Y that the causal significance of X for Y does not explain. Most of the correlation is explained not by discrimination, but, again, by the tendency of women to apply to the more stringent departments. See Figure 2.5.

All the examples so far are cases that show that correlation

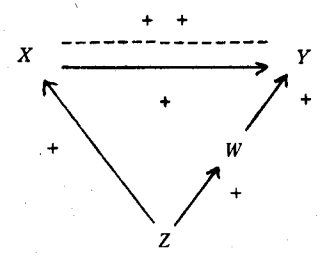


Figure 2.5

is no proof of causation. I mentioned above that it is also true that causation does not imply correlation. This can be seen easily enough by another modification of the Berkeley admissions example. If the males still tended to apply to the easy departments more frequently than the females did, but both departments discriminated, to just the right degree, against applicants believed to be male, then it could turn out that there is no overall correlation between getting admitted and being believed to be male. For an example of this, change the top row of the table above describing the original Berkeley example to read, "26/90, 1/10, 27/100." In this case, there is discrimination against those believed to be male, but exactly 27 percent each of males and females get admitted. So X is causally negative for Y (positive for $\sim Y$), yet there is no correlation, overall, between X and Y (or between X and $\sim Y$). See Figure 2.6 (the 0 above the broken line between X and Y represents probabilistic independence, *no* correlation of Y with X). This we may call a case of "spurious independence."

The examples we have seen show that we need a more general understanding of the idea of spurious correlation than the one first given. One obvious generalization, to accommodate cases of spurious independence and some of the other cases of spurious correlation in which X *is* causally relevant to Y, is to say that Y is spuriously correlated with X if, because of the action of a separate cause of Y, it is not true that X is

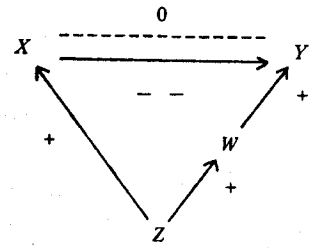


Figure 2.6

causally positive, negative, or neutral for Y, according to whether $Pr(Y/X)$ is greater than, less than, or equal to, $Pr(Y/\sim X)$, respectively. Note that on this understanding, spurious *correlation* includes, somewhat awkwardly, the possibility of probabilistic *independence* combined with some kind of causal relevance (spurious independence).

Note also that even though the relation of correlation itself is symmetric, I have used conditional probabilities only in one direction (from X and from $\sim X$ to Y, and not from Y or $\sim Y$ to X) in this characterization of spurious correlation. That is because the problem of spurious correlation can only be of real interest in one direction. Since at most one of any X and Y can precede the other in time, it follows from the requirement of temporal priority of causes (alluded to at the beginning of this chapter and to be clarified in Chapter 5) that, as far as the causal relation between X and Y is concerned, there is just one question of interest, namely, "What is the causal role of the earlier factor for the later?" And we have already adopted the convention that X precedes Y. Because causes precede their effects, it seems needless to add to the reason why the later Y is not causally relevant to the earlier X any idea that X is only spuriously correlated with Y.

This characterization of spurious correlation is still not fully satisfactory. We have seen a case, intuitively a case of spurious correlation, that does not fit this description. In that

example (Figure 2.5), we may say that there are two "components" of the correlation of Y with X: a small component due to X's positive causal significance for Y, and a large component due to the existence of a common cause of X and Y. The fact of the second component makes the correlation of Y with X "largely spurious," or largely unrepresentative of the causal significance of X for Y. In this example, Y is strongly positively correlated with X, yet X is only weakly causally positive for Y, so that the *degree* of correlation of Y with X does not match, intuitively, the "degree of causal significance of X for Y." Although we can at this point be precise about the meaning of "degree of Y's correlation with X" (just the difference, $Pr(Y/X) - Pr(Y/\sim X)$), we cannot yet be precise about "degree of causal significance of X for Y." Yet I wish henceforth to understand spurious correlation more generally as follows: Y is spuriously correlated with X if, because of the fact of a separate cause of Y, the degree to which Y is correlated with Y does not equal the degree to which X is causally significant for Y.

Not being precise at this point about degree of causal significance is on a par with not having been precise about what causation is in the characterization earlier of the narrower idea of spurious correlation (as simply common cause correlation without causation). Just as the simpler idea of spurious correlation was characterized in terms of causation without being precise about causation, so also the more general idea is characterized in terms of degree of causal significance without being precise about degree of causal significance. After property-level probabilistic causation itself has been sufficiently clarified in the pages that follow, we will be in a position to be more precise about the meaning of degree of property level causal significance.[7]

The new, more general understanding of spurious correla-

[7]This will be done in the next section. It is perhaps worth noting that, as discussed more fully in that section, there will be a kind of circularity in the definitions, given there, of the various kinds of causal significance. The different kinds of causal significance a factor X can have for a factor Y will be defined in terms of *other causes* of Y. However, the definitions will *not* rely on the idea of *degree* of causal significance.

tion includes cases of spurious independence of Y of X, cases in which only part of the correlation between X and Y can be explained by X's causal relevance to Y, as well as cases in which negative causal relevance is accompanied by positive correlation, cases in which positive causal relevance is accompanied by negative correlation, and cases in which positive or negative correlation is accompanied by causal neutrality.

Note, incidentally, that the possibilities of positive correlation with causal neutrality, and of positive correlation with negative causal relevance, show that positive correlation is *not sufficient* for positive causal relevance; and the possibilities of positive causal relevance with probabilistic independence, and of positive causal relevance with negative correlation, show that positive correlation is *not necessary* for positive causal relevance. All four of these possibilities are illustrated either by an example given above, or by the result of changing the Y of an example given above to $\sim Y$, or vice versa. The fact of these possibilities (or, more formally, the fact that any kind of correlation can be reversed or made to disappear in subpopulations), is known as "Simpson's paradox," named for E. H. Simpson (1951).[8] Each of the possibilities of positive correlation, negative correlation, and probabilistic independence can be consistently combined with each of the possibilities of positive causal factorhood, negative causal factorhood, causal neutrality, and (what I will describe in the next section) mixed causal relevance.

So much for examples of (single separate cause) spurious correlation for now. It is time to see how they may be explained in general, so that the possibility of spurious correlation may be appropriately accommodated in the theory of probabilistic causation. Consider first the simple kind of spurious correlation in which Y is positively correlated with X and X is completely causally irrelevant to Y (such as in the

[8]Cartwright (1979) mentions that this is sometimes known as the Cohen–Nagel–Simpson paradox, since it is presented as an exercise in Cohen and Nagel (1934). She also says that Nagel suspects he learned about it from Yule's (1911).

Fisher smoking hypothesis example, the falling barometers and rainy days example, and the first version of the Berkeley admissions case discussed above). The crucial feature of this kind of spurious correlation, a feature that fully explains this kind of correlation, is that a genuine probability-increasing cause (Z) of Y is correlated with the noncause (X) of Y. Whether the correlation between the genuine cause and the noncause is spurious or not is irrelevant to whether or not the correlation between X and Y is spurious. The point is that when the noncause (X) occurs, *the genuine cause (Z) is simply more likely to occur,* thus increasing the probability of Y.

It is because X is correlated with a genuine, probability-increasing cause of Y, that X increases the probability of Y. And X's correlation with a genuine, probability-increasing cause of Y will result in a spurious correlation between X and Y whether or not the correlation between X and the genuine cause is spurious. And it is an entirely different question whether or not there will always be, in cases in which X is correlated with a genuine cause of Y, a genuine cause of Y that is also a genuine cause of X.

In the first, simple common cause examples discussed above, the genetic condition is a genuine cause of lung cancer and it is correlated with (because it causes) the noncause, smoking; and the passing of a cold front is a genuine cause of rain and it is correlated with (because it causes) falling barometers. In the Berkeley admissions example, the spurious correlation between being believed to be male and getting admitted is explained by the correlation between applying to a lenient department (a genuine cause of admission) and being believed to be male (the noncause) – even though this latter correlation is spurious. Of course, the spurious correlation is also explained by the correlation between being believed to be male and *being male,* where the latter factor is a genuine cause of admission as well as of being believed to be a male.

Consider now other kinds of spurious correlation between factors X and Y, cases in which X may be genuinely causally

relevant to Y but in which the degree of (overall) correlation between X and Y does not appropriately reflect X's true causal significance for Y. In these cases, it is incorrect to refer to X as a "noncause" of Y, as in the diagnosis given above of the simpler kind of spurious correlation. But the same diagnosis applies. The fact that X is not a noncause of Y does not affect the fact that the spurious correlation (the inequality between the correlation of Y with X and the degree of causal significance of X for Y) is explained by the existence of a factor Z that is correlated with X and is a genuine probability-increasing cause of Y. It is easy to see that this is the explanation for the spurious correlations in the other versions of the Berkeley admissions case considered above and in Skyrms's example involving smoking in the cities and in the country. In all these cases, there is a "component" of the correlation of Y with X that is explained not by X's causal significance for Y, but rather by the correlation, with X, of a separate genuine, probability-increasing cause of Y.

It is important to note, however, that not all cases in which a factor X is correlated with a genuine cause Z of a factor Y are cases of spurious correlation of Y with X. For example, in some cases of transitive causal chains from X to Z to Y, X will be correlated with Z. If X is a genuine cause of Z and Z is a genuine cause of Y, then it can happen that Z, a genuine cause of Y, is correlated with X. Yet Y's correlation with X need not be spurious in such a case: if causation is *transitive* in this case, then X will be a genuine cause of Y.[9] And in such a case, the degree to which Y is correlated with X may exactly equal the degree of X's causal significance for Y, so that the correlation is not spurious. This, of course, is just a reiteration of the reason, given earlier, for explicitly excluding, from cases of spurious correlation, cases in which the reason why a third factor Z screens off a correlation between factors X and Y is that Z is causally intermediate between X and Y.

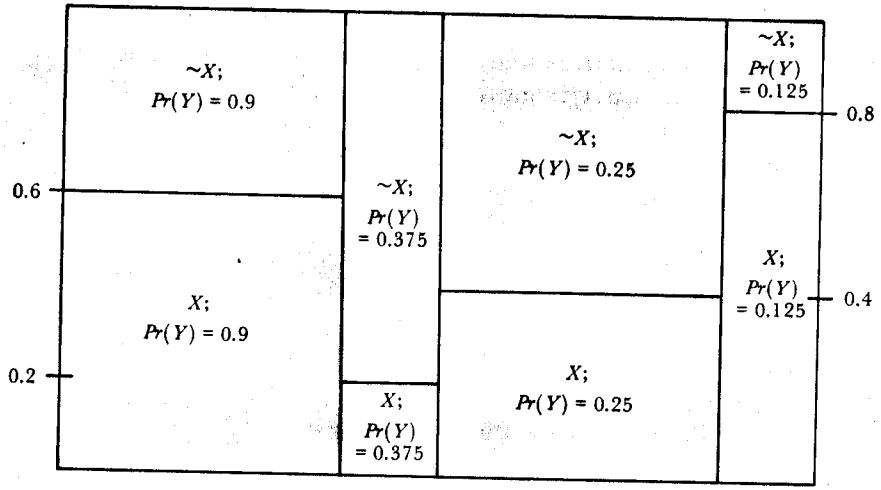
[9] Transivity of causal chains will be discussed in detail in Section 4.3.

What allows for the possibility of a spurious correlation between two factors X and Y is the existence of a third factor Z that is causally relevant to Y *independently of X's causal role, if any, for Y*. In all of our examples, the spurious correlation of Y with X is explained by the existence of a factor Z such that (i) Z is genuinely causally relevant to Y, (ii) Z is correlated with X, and (iii) X is causally irrelevant to Z. It is exactly this that made it possible, in the examples given above, for the degree of correlation of Y with X to fail to reflect just the causal significance of X for Y. In all the examples above, the factors Z (and, where applicable, both Z and W) satisfy the three conditions just laid down.

In cases of spurious correlation in which there is *just one* separate cause of the later factor Y (or in which other causes W trace back to a single separate cause Z), it seems that it is exactly (i), (ii), and (iii) above that explain the spurious correlation. These we may call cases of "single separate cause spurious correlation." When there are *multiple* separate causes, however, this diagnosis is not quite on the mark. It is possible for a factor Y to be spuriously correlated with a factor X and for this to be explained by the operation of, for example, *two* separate causes, F and G, of Y, *where neither F nor G is correlated, overall, with X*. In such a case, (i) and (iii) above hold of each of F and G, but (ii) fails of each of F and G. However, as we shall see, a condition very much like (ii), but involving *conditional* correlations will hold in such cases.

Examples of this are more complex and harder to grasp intuitively than examples of single separate cause spurious correlations.[10] Figure 2.7 depicts a numerical example of this kind; Figure 2.7a gives the probabilistic relations and Figure 2.7b gives the causal structure of the example. Although I will not describe a "real life" example corresponding to Figure 2.7, I would encourage the reader, after finishing this

[10] I note that Fisher and Patil (1974) give an example of this, and diagnosis it in terms of conditional correlations. Compare also Miettinen (1970) and (1974).



	F&G	*F&~G*	*~F&G*	*~F&~G*
$Pr(\pm F\&\pm G)$	=0.375 (3/8)	0.125 (1/8)	0.375 (3/8)	0.125 (1/8)
$Pr(X)$ =	0.6	0.2	0.4	0.8
$Pr(Y/X)$ =				
$Pr(Y/\sim X)$ =	0.9	0.375 (3/8)	0.25 (2/8)	0.125 (1/8)

OVERALL: $Pr(F/X) = Pr(F/\sim X) = 0.5$; $Pr(G/X) = Pr(G/\sim X) = 0.75$; so that *each of F and G is uncorrelated with X* - and each of *F* and *G fail to satisfy (ii)* above, though of course they *can still satisfy (iii)* (see Figure 7b below).

OVERALL: $Pr(Y/F) = 0.70625 > Pr(Y/\sim F) = 0.10625$, and $Pr(Y/G) = 0.575 > Pr(Y/\sim G) = 0.25$ (and the same *inequalities* hold when *G* and *F*, respectively, are held fixed, positively and negatively), so that *each of F and G is a cause of Y* - each of *F* and *G do satisfy (i)* above.

OVERALL: $Pr(Y/X) = 0.52375 > Pr(Y) = 0.49375 > Pr(Y/\sim X) = 0.46375$; so that there is a (*spurious*) *correlation of Y with X.*

Figure 2.7a

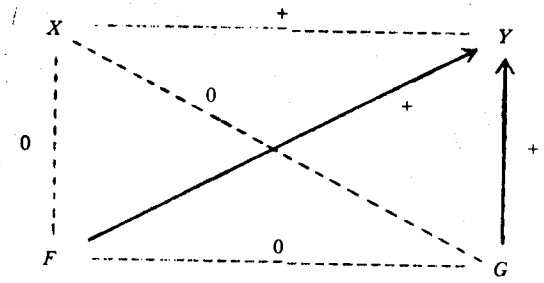


Figure 2.7b

section, to try to think of one. (I should add that there would not be a great loss of continuity if the reader were to skip to the end of this section at this point.)

Here, what explains the spurious correlation of Y with X is *not* the existence of any independent (of X) cause of Y that is, overall, *correlated with* X.[11] One way of explaining the correlation of Y with X in this example would be to note that, when the cause G is operating (which has probability 0.75), there *is* a correlation of the *stronger* cause F with X: A little calculation (or contemplation of Figure 2.7a) shows that $Pr(F/G\&X) = 0.6 > Pr(F/G\&\sim X) = 0.4$. There is this *conditional correlation* of F with X. When G is present, which is more probable than its being absent, X is correlated with the stronger cause, F, of Y. The *overall independence* of F from X is then explained by the fact that, when G is absent, there is a stronger negative correlation of F with X: $Pr(F/\sim G\&X) = 0.2 << Pr(F/\sim G\&\sim X) = 0.8$. These two conditional correlations, in opposite directions, are consistent with X's raising the probability of Y, overall, because G, within which there is a positive correlation between X and F, is so much more probable than $\sim G$, within which there

[11] The idea that there may be a *conjunctive* factor, such as $F\&G$ or $\sim F\&\sim G$, that is an independent cause of Y, and indeed correlated with X, will be discussed below.

is strong negative correlation between X and F, and because F is a stronger cause of Y than G is.

This suggests that spurious correlations should be explained, in general, in terms not only of *correlated independent causes,* but also of *conditionally correlated independent causes,* conditional on other causes – where, as we have seen, these conditional correlations are consistent with overall independence. That is, we could explain a spurious correlation of a factor Y with a factor X as arising from the existence of factors $F_1, \ldots, F_n$, such that (i) the F_i's are causes (positive or negative) of Y, (ii) each F_i is either correlated (positively or negatively) with X *overall* or correlated (positively or negatively) with X *conditionally on some way of holding fixed other* F_j*'s,* and, of course, (iii) the F_i's are causally independent of X. Without trying to prove this diagnosis rigorously here, I conjecture that this is the form of the most general kind of explanation of spurious correlation, where there may be multiple independent causes of Y.[12]

Before closing this section, there are two points I should make, both of which relate what has been discussed in this section to ideas that will be addressed later. First, about the example depicted in Figure 2.7, it might be suggested that, if we are clever enough, we *could* characterize, in terms of F and G, items that are causes of Y, independent of X, and that *are correlated with* X. For example, it might be suggested that some of the conjunctive factors, $F\&G$, $F\&{\sim}G$, ${\sim}F\&G$, and ${\sim}F\&{\sim}G$, may be independent causes of Y that are correlated with X. Or it might be suggested that we adopt a more general framework in which we consider *partitions* of factors (such as the *four-element partition* of consistent conjunctions

[12] However, it is easy to demonstrate this: If (1) there is no overall or conditional correlation between X and independent causes of Y (conditional on ways of holding other such causes fixed), and (2) maximal ways of holding fixed the independent causes screen off Y from X (which we would expect if X is not a cause of Y), then there can be no correlation of Y with X.

of F, G and their negations),[13] rather than simple "On/Off" (or *two-valued*) factors (such as F and G), as the relevant items that enter into causal relations. Note that the conditional probabilities of $F\&G$, $F\&{\sim}G$, ${\sim}F\&G$, and ${\sim}F\&{\sim}G$, all conditional on X, are all different in the example above, so that there is a kind of "correlation of the partition with X." These are approaches to spurious correlation that I would like to avoid.

As to the first suggestion, the *absence* of a conjunctive factor is disjunctive; for example, the negation of $F\&G$ is $(F\&{\sim}G) \vee ({\sim}F\&G) \vee ({\sim}F\&{\sim}G)$. And often the different disjuncts of a disjunctive causal factor confer different causally significant probabilities on the effect factor in question. So the question arises of how to average or otherwise combine these probabilities to come up with a *single* probability of the effect factor in the absence of the conjunctive factor, to compare with the probability of the effect factor in the presence of the conjunctive factor. Disjunctive causal factors are discussed in Chapter 3; the question just posed will be answered there in a way that makes the first suggestion ill suited as an approach to spurious correlation.[14]

As to the second suggestion, involving partitions, I think this would be just fine as an alternative approach to structuring our understanding of probabilistic causation. This makes the *form* of probabilistic causal claims quite different from the

[13] A *partition* is a set of factors (or propositions or sentences) that are "mutually exclusive" and "collectively exhaustive." That is, the conjunction of any two of its elements is impossible and the disjunction of them all is necessary. Appendix 1 explains the idea of a partition.

[14] Saying this, however, may help at this point. It may be argued that, in the example, $F\&G$ and ${\sim}F\&{\sim}G$ are both correlated with X and causes (positive and negative, respectively) of Y. They are causes of Y because no matter how we average the conditional probabilities of Y, conditional on the disjuncts comprising the absence of these conjunctive factors, the inequality between the probability of Y in the presence of one of the conjunctions and the probability of Y in its absence cannot change in direction. The problem, to be addressed in Chapter 3, is that the same is not true of $F\&{\sim}G$ and ${\sim}F\&G$. This is because these factors confer neither the highest nor the lowest probability on Y.

way I have been supposing we may fruitfully understand it. I have been supposing that we may understand the form of such claims to be (roughly), "X is a causal factor for Y in population P (considered to be of a kind Q)," where X and Y are "On/Off" (or two-valued) variables, or factors. I see no compelling reason to abandon this natural approach, and it is the approach I will continue to pursue.

Finally, it is worth emphasizing that the topic of this chapter is just spurious correlation, a phenomenon that involves *separate causes* of the candidate effect factor in question. As mentioned at the beginning of this chapter, there are other reasons, besides spurious correlation (that is, besides the confounding effects of separate causes), for why (degree of) correlation may not coincide with (degree of) causation. These involve *symmetry of correlation and asymmetry of causation* (to be discussed in Chapter 5) and *causal interaction* (to be discussed in Chapter 3). It should be emphasized that the discussion in this chapter is not intended to handle *all* the reasons why probability change does not coincide with causation, but only the reason of separate causes.

Based on the understanding of spurious correlation given above, the second main qualification of the basic probability-increase idea of probabilistic causation will be intended to deal with this phenomenon. We must "control" for correlations (both conditional and unconditional) that may exist between a factor X, whose causal role we want to characterize, and other causes, Z, of the effect factor Y in question. This is the topic of the next section.

2.2 CAUSAL BACKGROUND CONTEXTS

Recall the example about falling barometers (X) and rainy days (Y). Falling barometers do not cause rainy days, though rainy days are correlated with falling barometers: $Pr(Y/X) > Pr(Y/\sim X)$. Also, the factor of an approaching cold front (call it factor F here) is causally positive for rain, and it is correlated with (since it is a cause of) falling barometers. As we saw in the previous section, it is the existence of such factors as F here that explains the spurious correlation of Y with X.

Let us suppose that we know that *F is the only factor that is causally relevant to Y*, so that we have a case of what I called, in the previous section, "single separate cause spurious correlation." (This means, of course, that we know also that *X is causally neutral for Y*.) If we observe the probabilistic impact of X on Y first in the presence of F, and then separately in the absence of F, then we should expect, in each case, that Y will be probabilistically independent of X:

$$Pr(Y/F\&X) = Pr(Y/F\&\sim X)$$

and

$$Pr(Y/\sim F\&X) = Pr(Y/\sim F\&\sim X).$$

This is because we know that the *only reason* for a correlation of Y with X could be a correlation of F, a genuine cause of Y, with X; we know that X itself is causally neutral for Y.[15] And in the above two probability comparisons, F is "held fixed," positively and then negatively; and given each way of holding F fixed, the correlation of F with X disappears. When F is held fixed positively, the probability of F is 1, both conditional on X and conditional on $\sim X$; and when F is held fixed negatively, the probability of F is 0, both conditional on X and conditional on $\sim X$. Since we know that the *only reason* for a correlation of Y with X is the correlation of F, a genuine cause of Y, with X, we should expect any correlation of Y with X to disappear when the correlation of F with X disappears.

Suppose, on the other hand, that all we know, besides the *correlation of Y with X*, were the *two probabilistic equalities displayed above*, and that *aside from the possibility of X, F is the only factor causally relevant to Y*. Then I claim we should be in a position to conclude that X is causally neutral for Y. We

[15] I am, as always, following the convention that X precedes Y in time, as explained earlier.

should be able to conclude that Y's positive overall correlation with X is *explained entirely by a correlation of X with F*, the "single separate cause" of Y. In fact, in this example, we shall be in a position to apply Reichenbach's principle of the common cause involving conjunctive forks, as discussed at the beginning of the previous section. The problem is to explain the correlation of Y with X. The equalities above, which I have just assumed we know to hold, are parts (3) and (4) of a conjunctive fork (substituting "F" for "Z" in the description of conjunctive forks earlier). Explaining why we would observe inequalities (1) ($Pr(X/F) > Pr(X/\sim F)$) and (2) ($Pr(Y/F) > Pr(Y/\sim F)$) in this example is a little more intricate.

First, why should we observe inequality (2) to hold? I am assuming that it is true that (and that we know that) aside from the possibility of X, F is the *only* factor that is causally relevant to Y. Now *in fact,* X is *not* causally relevant to Y (though this is what we will be reasoning *to,* and not *from*). Given this, and the fact that aside from the possibility of X, F is the only cause of Y, it follows that F in fact *is the only cause of* Y (though we cannot reason to this conclusion, just from the equalities above and the rest of what I have assumed, two paragraphs back, that we know). And this implies, I assume, that in fact, F raises the probability of Y overall. F cannot be correlated with *negative* causes of Y, resulting in a *spurious independence* of Y from F, or a *spurious negative correlation* between Y and F, for example: given *what we know to be true,* and given *the fact that* X actually is not a cause of Y, there cannot be any such negative causes of Y. That is why we should *observe* inequality (2) to hold (though we cannot *reason* to it just from what I have assumed, two paragraphs back, that we know). As to inequality (1), it *mathematically follows from* (2), (3), (4), and the overall correlation of Y with X.[16]

[16] From the two equalities above, it follows that:

$$Pr(Y/X) = Pr(F/X)Pr(Y/F) + Pr(\sim F/X)Pr(Y/\sim F).$$

and,

$$Pr(Y/\sim X) = Pr(F/\sim X)Pr(Y/F) + Pr(\sim F/\sim X)Pr(Y/\sim F).$$

Let $x = Pr(F/X)$ and $y = Pr(F/\sim X)$. We want to show that $x > y$. Let $a = Pr(Y/\sim F)$. By the assumption that F is the only cause of Y, and hence that Y is positively correlated with F, there is a positive number e such that $Pr(Y/F) = a + e$. Then the correlation of Y with X can be expressed as follows:

$$x(a + e) + (1 - x)a > y(a + e) + (1 - y)a.$$

which reduces to $xa + xe + a - xa > ya + ye + a - ya$, and then to $xe > ye$, and then, since $e > 0$, to $x > y$.

Thus, F, X, and Y form a conjunctive fork. This means that the correlation of Y with X is explained entirely by the existence of the genuine cause, F, of Y, and by F's correlation with X (and by the fact that there are no other separate causes of Y). Thus, we should conclude that the correlation of Y with X does not correspond to causal relevance, and in fact the disappearance of the correlation, when the single correlated cause is held fixed, indicates causal neutrality. From the equalities, and the assumption that, aside from the possibility of X, F is the only factor causally relevant to Y, we conclude that none of X's positive probabilistic relevance to Y could be explained by any *causal* relevance of X to Y. That is, given the assumption that we have held fixed *all* the causes of Y (aside from the possibility of X), we can conclude from *X's probabilistic neutrality for Y* that *X is causally neutral* for Y.

It is worth noting that all that is crucial in the reasoning, above, first, from X's causal neutrality for Y to the two displayed probabilistic equalities, and second, from the two equalities to causal neutrality, is this: The assumption that we know that, *aside from the possibility of X, F is the only factor that is causally relevant to Y.* It is not crucial that we know in addition that there is a correlation of Y with X. This is already more or less explicit for the first line of reasoning, from causal neutrality to the two equalities. For the second line of reasoning, from the two equalities to causal neutrality, it is easy to see that the reasoning can easily be modified so that the crucial assumpiton gives this conclusion: If the two equalities hold, then any correlation that there may be between X and Y would be fully explained by a correlation of F with X. That is, no component of any correlation between X and Y is explainable by any causal relevance of X to Y. Given the basic probability-

increase idea for probabilistic causation, this means that we can conclude, from the inequalities, that X is causally neutral for Y. (Note also that, given the two equalities and Y's correlation with F, Y will be correlated with X if and only if F is correlated with X.)

Thus, the main point is this: In evaluating the causal role of a factor X for a factor Y, if a factor F is the only factor that, aside from the possible causal role of X, is causally relevant to Y, then, when we hold F fixed positively and negatively, probabilistic independence of Y from X coincides with causal neutrality of X for Y – whether or not Y is, overall, correlated with X. After explaining how this idea deals with spurious correlation more generally (including spurious independence and spurious degrees of correlation, as well as cases of multiple separate causes), I will show how the resulting theory deals with some of the other examples discussed in the previous section.

The basic idea behind this refinement of the theory is that in order for positive, negative, and neutral causal relevance to coincide with positive, negative, and neutral probabilistic relevance, we have to control for all factors that are, independently of the candidate causal factor in question, causally relevant to the effect factor in question. In order for X to count as a positive causal factor for Y, for example, X must have a positive probabilistic impact on Y beyond that which is explainable by *other,* independent, causes of Y that may be correlated with X (where, as noted at the end of the previous section, this correlation may be either unconditional or conditional on other causes of Y). The general strategy for controlling for independent causes goes back a long way; Skyrms (1980) reports that the basic idea was already explicitly formulated by F. Y. Edgeworth (1892, 1910), was anticipated half a century earlier by Bravais, and was used by the English statistical school of Edgeworth, K. Pearson (1897), and G. U. Yule (1910). See also, for example, Reichenbach (1956), H. Simon (1957), C. Granger (1969), Suppes (1970, 1984), Salmon

(1971), Cartwright (1979), Skyrms (1980, 1984a), Eells and Sober (1983), and, for a review, Skyrms (1988).

Suppose we want to assess the causal significance of a factor X for a factor Y. And suppose there are exactly n factors, $F_1, \ldots, F_n$, distinct from X, that are causally relevant to Y in a way that is independent of X's causal relevance, if any, to Y. That is, the F_i's are all the factors, other than X and effects of X, that are causally relevant to Y.[17] We saw in the previous section that it is the existence of such factors F_i *that are correlated with* X (either overall or conditional on other such factors) that makes possible a spurious correlation between X and Y. However, in the theory described below, causes of Y that are *uncorrelated* with X (either overall or conditional on other such causes) will be controlled for in the same way as causes of Y that *are correlated* with X (either overall or conditional on other such causes). As explained below, first, it *cannot hurt* to do so, in that whether or not we control for uncorrelated causes, we get the same answers about both direction and magnitude of causal influence; and second, this approach will give "causal background contexts" within which it can only be exactly the causal impact *of* X on Y (as opposed to the causal impact of *other causes* as well) that is measured by the probabilistic impact of X on Y (the magnitude of probabilistic impact within contexts will not be skewed in ways in which it could be if we did not control for uncorrelated causes).[18]

Since the F_i's are *all* the factors that are (independently of X) causally relevant to Y, if we "control" for the presence or absence of each F_i, we can observe the probabilistic impact that X makes upon Y beyond that made by other factors that

[17] Again, I will postpone a detailed discussion of the important *independence* requirement until Chapter 4. The idea is that the F_i's should not include factors to which X is causally relevant – and that, as explained below, we should not "hold fixed" causally intermediate factors when assessing the causal role of X for Y.

[18] I am hedging here: We will see in Chapter 3 that there are sometimes factors that are *not causes* of the candidate effect factor, but which nevertheless must be controlled for, namely, "interactive" factors.

are causally relevant to Y (independently of X). If we observe that X increases the probability of Y no matter what other independently causally relevant factors are present, then (pending further refinements of the theory that do not involve spurious correlation) we can conclude that X is causally positive for Y.

In assessing X's causal relevance to Y, we have to hold fixed all the other factors, $F_1, \ldots, F_n$, that are causally relevant to Y, independently of X, and then observe the probabilistic impact of X on Y given each of these ways. With n such other factors, there are 2^n ways of holding each fixed, positively or negatively. That is, there are 2^n conjunctions in which each of these n factors occurs exactly once, either positively (unnegated) or negatively (negated). Of these 2^n "maximal conjunctions," let $K_1, \ldots, K_m$ be exactly those that have nonzero probability both in conjunction with X and in conjunction with $\sim X$. (Thus, for $i = 1, \ldots m$, $Pr(K_i\&X) > 0$ and $Pr(K_i\&\sim X) > 0$.) These K_i's are called "causal background contexts," relative to the assessment of X's causal role for Y.[19]

Then we say that X is a *positive causal factor* for Y if and only if, *for each i*, $Pr(Y/K_i\&X) > Pr(Y/K_i\&\sim X)$. *Negative causal factorhood* and *causal neutrality* are defined by changing the "always raises" ($>$) idea to "always lowers" ($<$) and "always leaves unchanged" ($=$), respectively. The idea that the inequality or equality must hold for *each* of the background contexts K_i is sometimes called the condition of *contextual unanimity*, or *context unanimity*.[20] This condition, and some alternatives to it, are discussed in the next section. Note that these three relations of positive, negative, and neutral causal factorhood

[19] If the F_i's already contain conjunctions of some of the F_i's and their negations, then it could turn out that m is less than n. Whether or not the F_i's do contain conjunctions of F_i's and their negations turns on the questions of the causal roles of disjunctive causal factors, which will be discussed in Chapter 3.

[20] The term "contextual unanimity" was introduced by John Dupré (1984) to distinguish this kind of "unanimity" from what Elliott Sober and I (1983) called "unanimity," by which we meant what Dupré calls "unanimity of intermediaries." I will discuss the latter idea in Chapter 4.

are not exhaustive of the possible causal significance that a factor X can have for a factor Y: There remains the possibility of various kinds of *mixed* causal relevance, corresponding to various ways in which unanimity can fail. This possibility is discussed further in the next section and in Chapter 3.[21]

Note further that the definition is circular in that it characterizes X's causal role for Y in terms of other causally relevant factors. However, the circularity is not as bad as it could be. As Skyrms (1988) has pointed out in connection with this kind of definition of causation, X's causal role for Y is characterized in terms of the causal roles of factors *other than* X for Y. The idea of X's being causally positive for Y, for example, is noncircularly defined in terms of the causal roles of *other* factors for Y. However, the definition of positive causal relevance in general *is* circular. Thus, what we have is a theory about the *relation* between probability and causation.[22]

Before applying these definitions to some of the examples of the previous section, four further points about the definitions are in order here. First, as stressed in Section 1.1, these definitions must be understood as relative to a particular population, as well as to a kind that the token population exemplifies. This relativity to a population and a kind is already explicit in the assumption that we have a definite probability function to work with: as explained in Section 1.1, probabilistic, as well as causal, relations can differ from population to population, and from kind of population to kind of population.

Second, it is worth reiterating here the important point briefly motivated in the previous section, and alluded to from time to time in this section, that we only hold fixed factors

[21] In particular, I will show in the next section that, among the factors $F_1, \ldots, F_n$ that are causally relevant to Y independently of X, and that must be held fixed in background contexts, we must include not only all factors that are causally *positive* or *negative* for Y independently of X, but also factors that are causally *mixed* for Y independently of X. *Mixed causal relevance is a kind of causal relevance.*

[22] In Chapter 5, I will consider a natural suggestion for removing the circularity (the idea is to hold fixed all factors simultaneous to or earlier than the causal factor in question). I will give an example that strongly suggests that we should reject this suggestion.

that are causally relevant to Y *independently of* X. As explained in the previous section, these are factors that are causally relevant to Y but to which X is *not* causally relevant. A simple rationale for this independence condition was, in effect, briefly explained in the previous section: X's correlation with factors that are causally relevant to Y, but *intermediate* in a causal chain from X to Y, may fail to make the correlation between X and Y spurious. However, the condition has been the subject of some controversy in recent literature on probabilistic causality. In Chapter 4, I discuss the condition in detail, explain various parts of the controversy, and defend and generalize the condition.

Third, as promised in the previous section, we can now be more precise about the idea of degree of causal significance that was used in characterizing spurious correlation. A natural definition, suggested by the definition above, is this: The *average degree of causal significance* of a factor X for a factor Y is given by

$$ADCS(X, Y) = \Sigma_i\, Pr(K_i)[Pr(Y/K_i\&X) - Pr(Y/K_i\&{\sim}X)],$$

where the K_i's are the causal background contexts appropriates for assessing X's causal role for Y (in the relevant population considered to be of the relevant kind).[23] The difference between this and magnitude of correlation, $Pr(Y/X) - Pr(Y/{\sim}X)$, is that $Pr(K_i)$, enters into the formula for average degree of causal significance *unconditional on* X, which is appropriate since X is causally irrelevant to the K_i's, which specify factors causally relevant to Y *independently of* X. Note, of course, that this is an *average,* and that while positive, negative, and neutral causal relevance imply that the average is positive, negative, and zero, respectively, the converse is not true. This is because of the possibility of mixed causal significance, for

[23] I note that I. J. Good (1961–2, 1983,1985) offers a quite different kind of definition of (what he calls) "the tendency of [X] to cause [Y]." This is discussed briefly in Section 5.3 below.

which the average could be any value strictly between -1 and $+1$.

Finally, as mentioned above, the F_i's in the definitions above are *all* the factors, other than X, that are, independently of X, causally relevant to Y – and not just those that are *also correlated with* X, either unconditionally or conditional on other such causes. But spurious correlation between factors X and Y was diagnosed in the previous section as resulting from there being factors that are both independently causally relevant to Y and correlated with X, either unconditionally or conditional on other such causes. So the question naturally arises of why it should be required that we hold fixed, in the background contexts, *all* factors independently causally relevant to Y, and not just those of the kind that have been implicated in the possibility of spurious correlation.

The first thing to note is that even if it were not required to hold fixed uncorrelated causes of Y, it cannot hurt to do so. Suppose, for example, that a factor X is a positive cause of Y, and that X raises the probability of Y for each way, J_i, of holding fixed independent causes of Y that *are correlated* with X, either unconditionally or conditional on other causes of Y. Now suppose that Z is an independent cause of Y that is uncorrelated with X, either unconditionally or conditional on other independent causes of Y. I first show that the average degree of causal significance of X for Y is the same, whether or not Z is held fixed in addition to the other independent causes.

Not holding Z fixed, we have

$$ADCS(X, Y) = \Sigma_i\, Pr(J_i)[Pr(Y/J_i\&X) - Pr(Y/J_i\&{\sim}X)].$$

Factoring in Z, this is the same as

$$\begin{aligned} ADCS(X, Y) = {} & \Sigma_i\, Pr(J_i) \\ & \{[Pr(Z/J_i\&X)Pr(Y/J_i\&Z\&X) + Pr({\sim}Z/J_i\&X)Pr(Y/ \\ & J_i\&{\sim}Z\&X)] - [Pr(Z/J_i\&{\sim}X)Pr(Y/J_i\&Z\&{\sim}X) + Pr({\sim}Z/ \\ & J_i\&{\sim}X)Pr(Y/J_i\&{\sim}Z\&{\sim}X)]\}. \end{aligned}$$

Since Z is uncorrelated with X, conditional on the J_i's, we have

$$ADCS(X,Y) = \Sigma_i\, Pr(J_i)$$
$$\{[Pr(Z/J_i)Pr(Y/J_i\&Z\&X) + Pr(\sim Z/J_i)Pr(Y/J_i\&\sim Z\&X)]$$
$$-[Pr(Z/J_i)Pr(Y/J_i\&Z\&\sim X) + Pr(\sim Z/J_i)Pr(Y/J_i\&\sim Z\&\sim X)]\}.$$

Then rearranging and simplifying, we have

$$\begin{aligned} ADCS(X,Y) &= \Sigma_i\, Pr(J_i)Pr(Z/J_i)[Pr(Y/J_i\&Z\&X) \\ &\quad - Pr(Y/J_i\&Z\&\sim X)] + \Sigma_i\, Pr(J_i)Pr(\sim Z/J_i)[Pr(Y/ \\ &\quad J_i\&\sim Z\&X) - Pr(Y/J_i\&\sim Z\&\sim X)] \\ &= \Sigma_i\, Pr(J_i\&Z)[Pr(Y/J_i\&Z\&X) - Pr(Y/ \\ &\quad J_i\&Z\&\sim X)] + \Sigma_i\, Pr(J_i\&\sim Z)[Pr(Y/ \\ &\quad J_i\&\sim Z\&X) - Pr(Y/J_i\&\sim Z\&\sim X)]. \end{aligned}$$

And this is $ADCS(X,Y)$ calculated in terms of background contexts $J_i\&Z$ and $J_i\&\sim Z$, which are obtained from the J_i's by holding fixed, in addition to the other independent causes of Y, Z as well. Thus, holding fixed factors like Z (causes of Y that are causally independent of X and uncorrelated with X overall and conditional on other independent causes of Y) does not affect the value of $ADCS(X,Y)$.

This agreement about *average degree* of causal significance does not by itself imply agreement on the *qualitative question* of kind of causal significance. However, if it suffices, for getting the right answer, to hold fixed just the causally independent (of X) causes of Y that are conditionally or unconditionally correlated with X (and the main point of previous section is that this does suffice, at least for dealing with spurious correlation), then, I claim, we should expect the same answer when we hold fixed such factors that are conditionally and unconditionally uncorrelated with X. That is, for contexts J_i and factors X, Y, and Z, as above, we should expect

$$Pr(Y/J_i\&X) > Pr(Y/J_i\&\sim X)$$

if and only if both

$$Pr(Y/J_i\&Z\&X) > Pr(Y/Y_i\&Z\&\sim X)$$

and

$$Pr(Y/J_i\&\sim Z\&X) > Pr(Y/J_i\&\sim Z\&\sim X);$$

and the same when we substitute "<" or "=" for ">" throughout. In fact, holding fixed *all* independent causes of Y (those held fixed in the J_i's as well as factors like Z above) gives us contexts within which causal impact is approximated by probabilistic impact more closely than if we only held fixed independent causes that are correlated, conditionally or unconditionally, with X, as I will now explain.

To simplify this discussion, let us suppose that there are no causes of Y that are correlated with X, conditionally or unconditionally, and that, aside from the possibility of X, Z is the only cause of Y; so Z is uncorrelated with X. (For cases in which there are causes of Y that are correlated with X, the possibility I describe below could arise within contexts J_i.) And let us suppose that X is a positive causal factor for Y. It is possible that X can make only either a very large difference for Y or a very small difference – and never a moderate, or intermediate, difference. And this feature of the causal significance of X for Y may not show up unless factors like Z are held fixed. For example, the relevant probabilities may be as follows:

$$Pr(Y/Z\&X) = 0.9 > Pr(Y/Z\&\sim X) = 0.2;$$
$$Pr(Y/\sim Z\&X) = 0.4 > Pr(Y/\sim Z\&\sim X) = 0.3;$$
$$Pr(Y/X) = 0.65 > Pr(Y/\sim X) = 0.25.$$

Here, $Pr(Z) = Pr(Z/X) = Pr(Z/\sim X) = 0.5$, so that the probability values in the last line displayed are 50-50 averages of the values in the first two lines.

In this example, X can either make a huge difference in the probability of Y, or only a tiny difference. When Z holds, the

difference is huge: $0.9 - 0.2 = 0.7$. When $\sim Z$ holds, the difference is tiny: $0.4 - 0.3 = 0.1$. And there is no kind of individual (or no kind of concrete situation) in which X can make a moderate, or intermediate, difference in the probability of Y. Nevertheless, on average (that is, not holding Z fixed), X does make a moderate, or intermediate, difference in the probability of Y: as the last line displayed above shows, this average difference is $0.65 - 0.25 = 0.4$. However, this *average difference in probabilities* does not correspond to a *degree of causal influence* that X can have on Y for any kind of individual, or in any concrete situation. For in any individual or concrete situation, Z will either be present or absent, so that the causal significance of X for Y will be either very large (0.7) or very small (0.1), and never intermediate (0.4).

By holding fixed *all* independent causes of Y, and not just those causes that are conditionally or unconditionally correlated with X, we can more accurately observe, in terms of probability comparisons, the causal impact that *just* X has on Y. By doing this, we control for the probabilistic significance that causes other than X can have for Y, even though these causes may not be correlated with X. This yields causal background contexts in which the causal significance of *just* X for Y is more precisely isolated. It is for this reason that the definitions above require holding fixed *all* independent causes of Y.

Let us now see how the definitions given above give the right answers in the examples of the previous section diagrammed in Figures 2.2–2.6. In the first, simple, version of the Berkeley admissions example, we have to hold fixed the factors Z (being male) and W (applying to a stringent department) in assessing the causal role of X (being believed by the school to be male) for Y (admission). It was part of the example that in stringent departments, as well as in departments that are not stringent, the frequency of admission is the same among those believed to be male as it is among those not believed to be male. So, holding fixed W (and holding fixed

Z as well should not affect this) makes Y probabilistically independent of X. So the definitions tell us that being believed to be male is causally neutral for admission, which is the right answer.

Figure 2.3 depicts the example in which being believed to be male is slightly causally negative for admission despite the positive correlation: Both stringent departments and departments that are not stringent tend to discriminate against applicants believed to be male. Again, the definitions tell us that we must hold fixed both Z and W, since they are causally relevant to Y independently of whatever causal relevance X has for Y. And the description of the example tells us that given each way of holding these factors fixed, the frequency of Y is less given X than it is given $\sim X$. So we get the right answer that being believed to be male is causally negative for admission. The reader is invited to apply the definition of average degree of causal significance to this example.

In Skyrms's example, diagrammed in Figure 2.4, we have to hold fixed the factor of living in the country (Z). And it is part of the example that both among the country dwellers and among the city dwellers, smoking decreases the probability of healthy lungs, so that the definitions give us the right answer that smoking is causally negative for pulmonary health.

Figure 2.5 diagrams the version of the Berkeley admissions example in which there is a little discrimination in favor of males both in the stringent and in the not so stringent departments, but a big correlation between getting admitted and being believed to be male, which is explained mainly by the common cause Z (being a male). Of course again we must hold fixed the factors Z and W, so that, in the relevant probability comparisons, the component of the overall correlation due to the common cause will disappear. That is,

$$Pr(Y/X) >\!\!> Pr(Y/\sim X),$$

but

$Pr(Y/Z\&W\&X) > Pr(Y/Z\&W\&\sim X)$ (by a little),
$Pr(Y/Z\&\sim W\&X) > Pr(Y/Z\&\sim W\&\sim X)$ (by a little),
$Pr(Y/\sim Z\&W\&X) > Pr(Y/\sim Z\&W\&\sim X)$ (by a little),

and

$Pr(Y/\sim Z\&\sim W\&X) > Pr(Y/\sim Z\&\sim W\&\sim X)$ (by a little).

The last four probability comparisons are the ones relevant to assessing the degree of X's causal significance for Y; and the differences in these are smaller than in the first comparison.

In the "causation without correlation" example of Figure 2.6, we again must hold fixed Z and W. And again the effect of holding fixed W reveals the policies of the stringent and the not so stringent departments: Within each kind of department, there is negative probabilistic relevance of being believed to be male and getting admitted, and therefore, according to the definitions, negative causal relevance of the former for the latter.

Thus, the definitions of the different kinds of causal factorhood given above provide plausible analyses of the examples of spurious correlation given in the previous section. Of course there are other ways in which we can have correlation without causation, aside from what we have been calling spurious correlation. There is still the problem of temporal asymmetry of causation and the problem of probabilistic causal interaction, both briefly described at the beginning of this chapter. In the next section, in the course of further evaluating the definitions given in this section, the problem of causal interaction will emerge. This will be dealt with by further refinements of the theory in Chapter 3. In Chapter 5, the problem of temporal priority of causes will be handled.

2.3 CONTEXT UNANIMITY

It has been questioned whether a genuine cause really must raise the probability of a genuine effect of it in *every* causal

background context. That is, the condition of context unanimity has been questioned. Skyrms (1980), for example, has suggested a weaker condition, which he calls a "Pareto-dominance condition": X raises the probability of Y in a least one causal background context ($Pr(Y/K_i\&X) > Pr(Y/K_i\&\sim X)$ for at least one i) and X lowers the probability of Y in no causal background context ($Pr(Y/K_i\&X) \geq Pr(Y/K_i\&\sim X)$ for every i). This can be called "Pareto-positive causal factorhood." The corresponding definition of negative causal factorhood, which can be called "Pareto-negative causal factorhood," would be parallel (in effect, X is Pareto causally negative for Y if it is Pareto-positive for $\sim Y$); and the definition of causal neutrality would remain the same. John Dupré (1984) has argued for a more radical departure from context unanimity. He proposes that the condition should be dropped altogether and replaced with an idea he calls "statistical correlation in a fair sample."

In this section, I will examine Skyrms' suggestion and Dupré's argument for rejecting context unanimity. I believe that neither revision is necessary, and that there are advantages in *not* revising the definitions in either of these ways. I will also argue that, while Skyrms's suggestion is fairly "harmless," Dupré's more radical departure from context unanimity is a step backward, one that must ultimately either make causation tantamount to mere correlation or, in order to avoid vagueness, involve arbitrary, unmotivated distinctions.

Skyrms offers no rationale for his suggested Pareto weakening of the original definition, simply calling it a "plausible interesting weakening" of the stronger condition. However, Elliott Sober (1984a) offers the following interesting rationale for the suggestion:

> Suppose some other physical condition, apart from smoking, *guarantees* the occurrence of a coronary. If an individual has that physical condition, smoking cannot boost the probability of a heart attack any higher that it already is. Yet it would be overly restrictive to conclude that smoking is not a positive causal factor for heart at-

tacks in a population that happens to include some individuals with the condition. To take account of this sort of case, we should relax the requirement in the following way: The causal factor must raise the probability of the effect in at least one background context and must not lower it in any. (pp. 293–4)

Although I can sympathize with the intuitions that motivate the weakening for Sober, and although I think the revision does not lead to any serious difficulties, it seems to me that there is a better way to account for cases like the one Sober describes.[24]

In this example, we can describe, using the non-Pareto (or strict) version of probabilistic causation, all the causal facts in the general population – by considering *subpopulations*. There are three relevant populations involved: (i) the general population, (ii) the subpopulation of individuals who lack that physical condition, and (iii) the small subpopulation of individuals who have that condition. According to the original definitions, smoking has a mixed causal role for heart attacks in the first population, a positive causal role for heart attacks in the second population, and a neutral causal role for heart attacks in the third.

An advantage to this approach is that it allows for a group of statements about causal relevance in particular populations to have greater descriptive power. Even when considering particular subpopulations, the Pareto-revision approach does not allow for expression of the idea that a causal factor is "unanimously positive" for the effect. Of course subpopulations can be found for which the Pareto-revision approach can express the truth that, within them, the causal factor is neutral for the effect factor. But if all we say about the rest of

[24] I should add that an important reason for Sober's using the Pareto formulation has to do with the strategy of his critique of genic and group selectionism in evolutionary theory (1984a). The Pareto version is *weaker* than the strict version in such a way as to give genic and group selectionism, characterized in terms of probabilistic causality, a "better chance" of being true; and if genic and group selectionism are false on the Pareto interpretation, then they must be false on the strict version as well.

the subpopulation is that the causal factor is Pareto-positive for the effect factor, then, consistent with this statement of Pareto-positive causal significance, the possibility remains that, in a subpopulation of *it*, the causal factor is neutral for the effect factor.[25]

In statements of *positive or negative* causal significance, it seems that we should want our concepts of positive, negative, and neutral causal factors to be just as sensitive to the possibility of causal *neutrality* within some contexts or subpopulations as they are to the possibility of the *reverse* kind of causal significance within some contexts or subpopulations. Intuitively, there are three "pure" (non-Pareto and unmixed) possible causal roles one factor can have for another: positive significance, negative significance, and causal neutrality. The Pareto condition allows mixtures of the first and the last to count as positive causal relevance. But a mixture of the first and the last is just as mixed a causal factor as a mixture of any other two of the three kinds of unmixed causal significance. Indeed, as will become clearer in Chapter 3, it seems best to think of cases of (nontrivial) Pareto probabilistic causation as examples of *causal interaction* (briefly explained at the beginning of this chapter), in which, due to the interaction, we have correlation without (strict) positive causal factorhood.

On the other hand, of course, one may carve up all the possibilities however one wants, and if I do it differently from the way you do it, then we simply arrive at *different concepts*. One set of concepts may be more versatile or descriptive than the other for one purpose, and vice versa for another purpose; and each set of concepts may be just as "legitimate" and coherent as the other. I do not think there is anything

[25] Of course if one adds the information that the first subpopulation is the *largest* subpopulation of the general population for which the causal factor is neutral for the effect factor, then we have as much causal truth described as is possible using the strict definitions. But a statement providing this kind of information goes beyond statements of the form "X is causally positive (or negative or neutral or mixed) for Y in populaiton P": such a statement *quantifies over* subpopulations.

conceptually wrong or incoherent with the Pareto-revision, of course. However, for the reasons given, I will stick with the strict understanding of positive and negative causal relevance, given in the original definitions of the previous section.

Let us now turn to the reasons advanced by Dupré (1984) for abandoning context unanimity altogether. The main consideration Dupré advances in support of abandoning the requirement is the possibility of the following kind of case:

> Suppose that scientists employed by the tobacco industry were to discover some rare physiological condition the beneficiaries of which were less likely to get lung cancer if they smoked than if they didn't. Contrary to what the orthodox [context unanimity] analysis implies, I do not think that they would thereby have discovered that smoking did not, after all, cause lung cancer. . . . If this is correct it seems to suggest that causes should be assessed in terms of average effect not only across different causal routes, but also across varying causal contexts. (p. 72)

There are three points I would like to make about cases of this kind, the first having to do with our understanding of "causal background contexts," the second having to do with how the definitions of the different kinds of probabilistic causal relevance can deal with examples of this kind (given a proper understanding of contexts), and the third with Dupré's suggestion of assessing causes in terms of their "average effect."

First, if we follow the understanding of contexts explained in the previous section, it is not at all clear that the theory requires us to hold fixed in the background contexts the rare physiological condition in Dupré's example. In the explanation of contexts given in the previous section as well as in, for example, Cartwright (1979), we are required to hold fixed all and only those factors (other than the causal factor in question and its effects) that are *themselves causally relevant* to the effect factor in question. But is having that rare physiological condition a cause – positive, negative, or even mixed – of lung cancer? It need not be, given the way Dupré has formulated his example.

Consistent with the description of the example, that physiological condition could be positive, negative, mixed, or neutral for lung cancer. (In Chapter 3, I explain in more detail these possibilities for factors like the rare physiological condition in Dupré's example.) The condition could be a strong positive cause of lung cancer; but for those with the condition, smoking helps a little. (The situation could be as depicted in Figure 3.3 of Chapter 3, where X is smoking, F is the physiological condition, and Y is lung cancer.) The condition could also be a strong negative cause of lung cancer, where the combination of the condition with smoking gives the best possible protection. (See Figure 3.2 with X, F, and Y interpreted as just explained.) Also, the condition could be mixed for lung cancer, where among those with the condition, whether or not one smokes makes a big difference, and among those without the condition, smoking makes little difference. In this case, the condition could be negative for lung cancer among smokers, and positive among nonsmokers. (See Figure 3.4 with X, F, and Y again interpreted as explained above.)

Finally, and most importantly, the condition could be *causally neutral* for lung cancer, if it is causally relevant to smoking in just the right way, so that smoking is causally intermediate between the condition and lung cancer. (Figure 3.5, discussed in Chapter 3, shows how this can happen.) In this case, the definitions given above say we *should not* hold fixed that rare physiological condition when assessing smoking's causal role for lung cancer.

Having said this, I nevertheless think that on a proper understanding of causal background contexts, the rare physiological condition in this example *should* be held fixed, in assessing smoking's role for lung cancer. This means that our understanding of contexts has to be revised, since on the current understanding we do not hold fixed any factors that are causally neutral for the effect factor in question. In Chapter 3, on interaction, we will see why we have to hold fixed some fac-

tors, like the physiological condition in Dupré's example, that may be causally *neutral* for the effect factor in question. (By the end of this section it will be apparent why it is necessary to hold fixed independent *mixed* causes of the effect factor in question.) Henceforth in this section, let us suppose, with Dupré, for the sake of discussion of his example, that the context unanimity theory *does* require us to hold the rare physiological condition fixed.

My second point about Dupré's example is simply to show how the strict context unanimity theory succeeds in capturing all the causal truth in the example. As in the discussion Skyrms's suggested Pareto weakening of the theory, we exploit the relativity of probabilistic causality to populations. In the subpopulation of individuals without that rare physiological condition, smoking is causally positive for lung cancer. In the subpopulation of individuals with the condition, smoking is a negative causal factor for lung cancer. And in the combined population, smoking has a mixed causal role for lung cancer.

Given the definitions of the previous section, the fact that smoking has *mixed* causal relevance for lung cancer in the combined population implies that, in the combined population, smoking is *not* a *positive* causal factor for lung cancer. (The definitions imply that *positive, negative, neutral,* and *mixed* causal relevance are mutually exclusive, as well as exhaustive, of the kinds of causal significance one factor can have for another in one population.) I agree that the claim that smoking is not a positive causal factor for smoking in the combined population *can* be *misleading* – especially if we put it, as Dupré does, as the claim that "smoking did not, after all, cause lung cancer." But I think it is misleading only to the extent that we lose sight of the population-relativity of probabilistic causation, and perhaps slip back into interpreting the causal claim in terms of the concept of *token* causation. The claim that *smoking is not a positive causal factor for lung cancer in population P* does *not* imply that there are no *subpopulations* of

P within which smoking *is* a positive causal factor for lung cancer, nor does it imply that there are no *individuals* in *P* for whom smoking is a token cause, or would be a token cause, of lung cancer.

When it is clearly seen that a claim of positive causal factorhood in a population *P* is quite a strong claim (involving context unanimity), so that the denial of positive causal relevance in the same population is a correspondingly weak claim, and when the question of population-level causal significance is properly untangled from questions about token level causal significance, then the denial of positive causal factorhood should no longer be misleading. Indeed, when all this is borne clearly in mind, it seems best to say that this is another example of the problem of *probabilistic causal interaction,* of probability increase due to interaction rather than to positive causal factorhood, as briefly described at the beginning of this chapter and discussed more fully in Chapter 3.

Perhaps many of us would still not wish to deny that smoking is a positive causal factor for lung cancer in Dupré's example. We may even wish to say that, in this example, smoking is a positive causal factor for lung cancer in the (overall) human population. And some may wish to say this even after the distinctions of the previous paragraph have been thoroughly digested. Perhaps our intuitive concepts are such that "*X* causes *Y* in population *P*" is judged to be true if in a significant subpopulation of *P*, *X* is a (context unanimous) positive causal factor for *Y*. Suppose (just to have an example) that this understanding perfectly matches our intuitions, that it is a perfectly coherent (though vague) concept, and that it is perfectly serviceable in all contexts in which we may ever actually wish to characterize population-level causal relations. This may all be so, but it does not mean that this way of describing the causal facts cannot be improved on. For example, the understanding of population causation in question is vague, involving the idea of a "significant" subpopu-

lation of *P*. In this respect at least, the context unanimity theory is an improvement.

In general, I think there are the following *two* kinds of criteria for the evaluation of philosophical theories – in particular, theories of probabilistic causation. First, a theory should be appropriately sensitive the ways in which we use the words denoting the concepts the theory is about (for example, the words, "positive causal factor," or "is a cause of"). Before the development of a philosophically adequate theory about something can begin, we must first obtain, from common or scientific usage of the relevant terminology or concepts, at least a rough idea or impression of the thing the theory is supposed to be a theory about. This is at least a typical starting point. But second, the theory should be sensitive also to philosophical standards such as: *avoiding vagueness, securing logical consistency, simplicity, non-"ad hoc-ness," expressive power* (the degree to which a variety of possibilities are describable in terms of ideas described and developed in the theory), and so on. My intention is to weigh heavily this second kind of criterion. In any case, in particular, the *vagueness* of the idea, mentioned in the previous paragraph, of a "significant" subpopulation, brings me to my third point about Dupré's argument.

That point is that there is the following problem for those who would reject the requirement of context unanimity, and would say, for example, that a factor *X* is causally positive for a factor *Y* when *X* raises the probability of *Y* in all but a *rare* causal context or subpopulation within which the probabilistic significance of *X* for *Y* may be reversed. In Dupré's example, that physiological condition is supposed to be "rare." Say that in the relevant combined population, 1 percent of all individuals have the condition, and that this counts as rare. But what if the condition were not so rare? What if 5 percent had it – or 15 or 25 percent, or 55, or 95, or 99 percent? If, for such possible populations, we continue to relax the condition of context unanimity, then it is clear that we would be revert-

ing to a "mere positive correlation" theory of probabilistic causation, which we have already seen ample reason to reject.

On the other hand, it seems that part of the intuitive rationale for *not* holding that condition fixed in the original example was that the condition was so rare. No reason for not holding fixed such a condition that is *not* rare was given. Indeed, it seems that our intuitions tell us that if that condition were not rare, but rather intermediate in frequency, then smoking would have a *mixed* causal role for lung cancer. So it seems that at some point in the progression of possible populations in which the condition becomes more and more frequent, we must begin to hold the condition fixed, so that smoking then becomes causally mixed for lung cancer. In addition, it seems that Dupré must also say that, if 99 percent of the relevant individuals have that condition, then smoking is causally negative for lung cancer, since this case is entirely symmetrical with the original case. Thus, later on in the progression, we must again relax the requirement of context unanimity and once again not hold the condition fixed. So the problem arises of specifying, and motivating, "cut-off" frequencies for the condition at which points we should begin and then cease to hold that physiological condition fixed. I cannot see how this can be done in a way that would not be arbitrary.[26]

On this kind of approach, the question of whether smoking is a population-level cause of lung cancer will turn on the population frequency of that physiological condition, and in an unacceptable way. Indeed, it seems that in this example, this question should turn not at all on the frequency of individuals with that condition. For example, a person contemplating be-

[26] Actually, Dupré does not advocate any such "cut-off" frequency approach, but rather an idea that probabilistic causation is correlation in a "fair sample" of the original population, a sample in which other causal factors are "fairly represented." His approach would seem not even to allow for the category of "mixed" causal significance in a population, which itself seems to be a step backward. Also, his conception of a "fair sample" is vague and problematic. I will not discuss this idea here; see Eells (1987a) for criticisms.

coming a smoker, and trying to assess the health risks, should not be so concerned with the population frequency of that condition, but with whether or not *he* has the condition. That is, the person should be concerned with which *subpopulation* he is a member of, the subpopulation of individuals with the condition (a population in which smoking is causally negative for lung cancer) or the subpopulation of individuals without the condition (a population in which smoking is causally positive for lung cancer). The population frequency of the condition can provide the decision maker with evidence about whether he is in a subpopulation in which smoking is causally positive, or in one in which it is causally negative, for lung cancer, but (except for the extreme frequencies of 0 and 1) it cannot settle the question, and hence cannot be definitive of whether he is in a population in which smoking is positive, or one in which it is negative, for lung cancer.

Even if cut-off frequencies *could* be properly motivated, and even if some other approach could be developed that is both in harmony with Dupré's intuitions and not tantamount to identifying causation with correlation, there still remains a further problem. If, in Dupré's example, a theory says that smoking is simply causally positive for lung cancer in the combined population, then statements of probabilistic causal connection, interpreted on that theory, would in many cases mask a significant causal truth: the fact that there is an "interaction" between the causal factor (smoking, in the example) and *other* causal factors (the rare physiological condition, in the example). Just as for the suggested Pareto weakening discussed before, statements of probabilistic causal connection interpreted on such a theory could not settle the question of whether or not there is such an interaction, of whether or not the causal significance of one factor for another varies from subpopulation to subpopulation. In Chapter 3, we will see how the definitions of the previous section must be revised in order to properly and generally accommodate the possibility of causal interaction.

There is another lesson to be learned from Dupré's example. So far, we have only seen examples that show why, when assessing the causal significance of a factor X for a factor Y, we must hold fixed factors that are, independently of X, either *positive* or *negative* causes of Y. Dupré's example suggests another example, one that shows why we must also hold fixed factors that are, independently of X, causally *mixed* for Y.

Suppose things are the way Dupré describes them in his example, and suppose we are interested in the causal significance of *tobacco-stained fingers*, X, for lung cancer, Y, in the general population. Of course the truth is that the factor of tobacco-stained fingers is causally neutral for lung cancer. We saw above that smoking, in the example, is not a positive or negative cause of lung cancer, but rather mixed. And, as pointed out above, that rare physiological condition also need not be a positive or negative cause of lung cancer. But if we hold neither of them fixed, we can expect the probability of lung cancer given stained fingers to be greater than the probability of lung cancer given clean fingers: $Pr(Y/X) > Pr(Y/\sim X)$. This is because people who have stained fingers tend to be smokers and because that physiological condition is so rare in the general population. So if the theory told us to hold neither fixed, it would give the wrong answer that tobacco-stained fingers is a positive causal factor for lung cancer.

Now suppose we hold fixed the factor of that rare physiological condition, F (as already mentioned, the theory will be revised in Chapter 3 to require this). Then we can expect that stained fingers would decrease the probability of lung cancer among those with the condition and increase that probability among those without the condition: $Pr(Y/F\&X) < Pr(Y/F\&\sim X)$ and $Pr(Y/\sim F\&X) > Pr(Y/\sim F\&\sim X)$. This is because, among those who have the condition, stained fingers increases the probability that one is a smoker, which, among those with the condition, decreases the probability of lung cancer; and among those without the condition, stained fin-

gers again raises the probability that one is a smoker, which, among those lacking the condition, increases the probability of lung cancer. So if we hold fixed the physiological condition but not the factor of smoking, we get the wrong answer that stained fingers is causally mixed for lung cancer.

Recall, however, that the definitions given in the previous section say we should hold fixed *all* factors that are, independently of X, *causally relevant to* Y. If we interpret this as meaning, "all factors that have either positive, negative, *or mixed,* causal relevance to Y independently of X," then we must hold fixed the factor of smoking – call this factor G. This is because smoking has mixed causal relevance to lung cancer, independently of stained fingers. And, of course, given each of the four ways of holding fixed both the physiological condition *and smoking,* the correlation between stained fingers and lung cancer disappears:

$$\begin{aligned} Pr(Y/F\&G\&X) &= Pr(Y/F\&G\&\sim X), \\ Pr(Y/F\&\sim G\&X) &= Pr(Y/F\&\sim G\&\sim X), \\ Pr(Y/\sim F\&G\&X) &= Pr(Y/\sim F\&G\&\sim X), \end{aligned}$$

and,

$$Pr(Y/\sim F\&\sim G\&X) = Pr(Y/\sim F\&\sim G\&\sim X).$$

And this, of course, is simply because the only reason for the correlations (overall and conditional on F and on $\sim F$) between stained fingers and lung cancer is the correlation between stained fingers and smoking, and the causal roles of smoking for lung cancer among the F's and among the non-F's; so holding fixed, in addition to that physiological condition, the factor of smoking – which has mixed causal relevance to lung cancer independently of stained fingers – makes the correlation between stained fingers and lung cancer disappear.

Hence, we must interpret "causally relevant" – in the part of the definitions of positive, negative, mixed, and neutral causal factorhood that tells us what to hold fixed – as meaning "causally positive, negative, *or mixed.*" Also, incidentally,

the example just described is a case of a kind of spurious correlation we have not yet encountered. In this example, Y is spuriously correlated with X because X is correlated with (because caused by) a factor G that is, independently of X, a *mixed* cause of Y; and G is a genuine, probability increasing, common cause of both X and Y, *positive* for X and *mixed* for Y.

2.4 INTERACTIVE FORKS

According to our simplest understanding of spurious correlation, explained at the beginning of Section 2.1, two factors were spuriously correlated if neither causes the other, they are correlated effects of a common cause, and their correlation disappears when the common cause is held fixed. And recall that the probabilistic structure of such cases is given by

$$\begin{aligned} &(1)\ Pr(X/Z) > Pr(X/\sim Z), \\ &(2)\ Pr(Y/Z) > Pr(Y/\sim Z), \\ &(3)\ Pr(Y/Z\&X) = Pr(Y/Z\&\sim X), \\ &(4)\ Pr(Y/\sim Z\&X) = Pr(Y/\sim Z\&\sim X). \end{aligned}$$

Propositions (1)–(4) characterize common causes Z of factors X and Y in Reichenbach's (1956) sense, in which the presence, as well as the absence, of the common cause screens off the correlated effects from each other. Salmon (1978) calls this kind of probabilistic structure a "conjunctive fork."

However, there is another kind of probabilistic structure that has been recognized as a possibility for common cause situations. This is the kind of structure that Salmon (1978, 1984) has called an "interactive fork."[27] It is the same as a conjunctive fork except that one or both of (3) and (4) are changed to inequalities:

$$\begin{aligned} &(3\star)\ Pr(Y/Z\&X) > Pr(Y/Z\&\sim X), \\ &(4\star)\ Pr(Y/\sim Z\&X) > Pr(Y/\sim Z\&\sim X), \end{aligned}$$

[27] See also van Fraassen (1977b, 1980).

respectively. If a common cause situation has the probabilistic structure of an interactive fork, then holding fixed the common cause, Z, either positively or negatively, will fail to make its joint effects, X and Y, probabilistically independent each other. That is, one or both of Z and $\sim Z$ will fail to screen off Y from X.

Of course, this kind of possibility must be addressed in the theory of probabilistic causation; for according to the theory as developed so far, holding fixed independent (of X) causes Z of Y should render X probabilistically neutral for Y, if X is causally neutral for Y. There are several ways in which situations can exhibit the probabilistic structure of an interactive fork; that is, there are various *causal* patterns consistent with this kind of *probabilistic* structure. In order to properly assess the bearing of the possibility of interactive forks on the theory of probabilistic causation, we must be careful to distinguish among these.

We have actually already encountered several examples in which holding fixed a common cause fails to screen off its joint effects from each other. Figure 2.5, of Section 2.1, depicts one such example, an example that exactly fits conditions (1), (2), (3★), and (4★) for interactive forks. (Let us assume transitivity of the chain from Z to W to Y, so that Z is a probability-increasing cause of Y.) Recall that in the variation of the Berkeley admissions case depicted in Figure 2.5, Z (being male) is a probability-increasing cause of both X (being believed to be male) and Y (admission); and there is some discrimination in admissions policies in favor of males, so that X is, both in the presence and in the absence of Z, a probability-increasing cause of Y. So this example satisfies the conditions for interactive forks. But, of course, we have seen that the theory of probabilistic causation, as laid down so far, handles this case just fine. Since X *is* a cause of Y in the example (even though both are effects of a common cause), X *should*, according to the theory, increase the probability of Y, when we hold fixed the independent cause, Z, of Y.

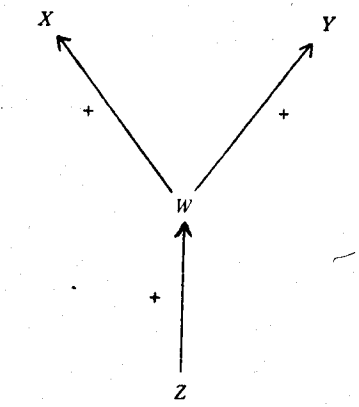


Figure 2.8

Another example having the probabilistic structure of an interactive fork is depicted in Figure 2.8. Here, Z, X and Y will form an interactive fork if these three conditions are met:

(5) W, X, and Y form a conjunctive fork, both conditional on Z and conditional on $\sim Z$,
(6) $Pr(W/Z) > Pr(W/\sim Z)$ (Z causes W),
(7) W screens off each of X and Y from Z, as does $\sim W$.

These three conditions imply that each of (1), (2), (3★), and (4★) will be satisfied by Z, X, and Y.[28]

[28] Essentially this was demonstrated in Eells and Sober (1986). Here is a somewhat different proof. For (1), note first that

$$Pr(X/Z) = Pr(W/Z)Pr(X/Z\&W) + Pr(\sim W/Z)Pr(X/Z\&\sim W)$$

and

$$Pr(X/\sim Z) = Pr(W/\sim Z)Pr(X/\sim Z\&W) + Pr(\sim W/\sim Z)Pr(X/\sim Z\&\sim W).$$

By (6), there are positive numbers a and u such that $a = Pr(W/\sim Z)$ and $a + u = Pr(W/Z)$. By (5), there are positive numbers b and v such that $b = Pr(X/Z\&\sim W)$ and $b + v = Pr(X/Z\&W)$. By (7), $Pr(X/\sim Z\&W) = b + v$ and $Pr(X/\sim Z\&\sim W) = b$. So, $Pr(X/Z) = (a + u)(b + v) + (1 - a - u)b = av + uv + b$, and $Pr(X/\sim Z) = a(b + v) + (1 - a)b = av + b$, from which (1) follows, since $uv > 0$. The proof of (2) is completely parallel. For (3★), note first that

$$Pr(Y/Z\&X) = Pr(W/Z\&X)Pr(Y/Z\&X\&W) + Pr(\sim W/Z\&X)Pr(Y/Z\&X\&\sim W)$$

and

Here is an intuitive example of this, a variation of the example given earlier about cold fronts, falling barometers, and rainy days.[29] Let X be barometers falling, Y be rainy days, and W be approaching cold fronts. And suppose that W, X, and Y form a conjunctive fork as before (only now we are using "W," instead of "Z," for approaching cold fronts). Now suppose that a cause of approaching cold fronts, in this part of the world, is westerly winds; call this factor Z. Then Z is a probabilistic cause of W, which in turn is a probabilistic cause of each of X and Y. By transitivity, which is plausible in this case, westerly winds is a probability-increasing cause of both falling barometers and rainy days, with approaching cold fronts as the intermediate causal factor. So (1) and (2) should hold in this example.

Now suppose that there is a westerly wind; that is, let us hold Z fixed positively. This does not necessitate the approach of a cold front. If we now add the information that barometers are falling, this should increase the probability that a cold front is approaching, which in turn increases the probability of rain. And if we instead added the information that the barometers are *not* falling, then this should decrease the probability that a cold front is approaching, which in turn decreases the probability that it will rain. So (3*) should hold. Now suppose that there is no westerly wind; that is, now hold Z fixed negatively. This does not necessitate there being no cold front approaching. So additional information to the effect that the barometers are or are not falling, increases and decreases, respectively, the probability that a cold front is

$$Pr(Y/Z\&\sim X) = Pr(W/Z\&\sim X)Pr(Y/Z\&\sim X\&W) + Pr(\sim W/Z\&\sim X)Pr(Y/Z\&\sim X\&\sim W).$$

By (6) and symmetry of correlation, there are positive numbers a and u such that $a = Pr(W/Z\&\sim X)$ and $a + u = Pr(W/Z\&X)$. Also by (6), $Pr(Y/Z\&X\&W) = Pr(Y/Z\&\sim X\&W) = Pr(Y/Z\&W)$, and $Pr(Y/Z\&X\&\sim W) = Pr(Y/Z\&\sim X\&\sim W) = Pr(Y/Z\&\sim W)$. And by (6) again, there are positive numbers b and v such that $b + v = Pr(Y/Z\&W)$ and $b = Pr(Y/Z\&\sim W)$. So, $Pr(Y/Z\&X) = (a + u)(b + v) + (1 - a - u)b = av + uv + b$, and $Pr(Y/Z\&\sim X) = a(b + v) + (1 - a)b = av + b$, from which (3*) follows. The proof of (4*) is completely parallel.

[29] For another example, see Section 4.2 (the indeterministic version of the example depicted in Figure 4.5).

approaching, which in turn increases and decreases, respectively, the probability of rain. And this is (4*).

In this example, Z is a common cause of X and Y, X and Y are causally neutral for each other; yet the common cause Z fails to screen off the correlation between X and Y. However, this kind of situation poses no problem for the theory of probabilistic causation laid down so far. It is not presupposed by the theory that in all cases in which joint effects of a common cause are causally neutral for each other, the common cause must screen off the effects from each other. It is only presupposed, so far, that, in such cases, when *all* independent causes of one of the joint effects are held fixed in one way, the other effect must be probabilistically neutral for the first. The joint effects must be independent conditional on specifications of *all* the causes of one of them (that are causally independent of the other).

In our example, each of W and Z is causally relevant (independently of each of X and Y) to each of X and Y. So, according to the theory, to assess the causal role of X for Y (or of Y for X) we must compare the probability of Y given X to the probability of Y given $\sim X$ (or X given Y and X given $\sim Y$) conditional on each of the four ways of holding fixed both W and Z. And condition (5) of the example implies that X and Y are independent conditional on each of the four ways, $W\&Z$, $W\&\sim Z$, $\sim W\&Z$, and $\sim W\&\sim Z$, of holding fixed W and Z.

This example shows that in at least *some* cases of interactive forks in which the joint effects are causally neutral for each other, when a finer description of the case is made, a conjunctive fork, and a screening off common cause, can be recovered. It is this feature of such cases that allows the probabilistic theory of causation to deliver the correct answers about what causes what in these cases. Also, in the example just discussed, the screening off common cause, W, is a factor that occurs *after* the time of the nonscreening common cause, Z (and, of course, before the time of the joint effects). These

two features of this example are shared by examples of interactive forks that have been discussed recently in the philosophical literature, with one important exception that will be discussed below. To illustrate this, and to demonstrate the versatility of the probabilistic theory, I will now turn to some recently discussed examples, and finally to the exception.

Salmon (1984) describes the following example. There are two balls on a pool table, the cue ball and the 8-ball. They are so situated that if a certain novice player attempts to put the 8-ball into a far corner pocket by shooting the cue ball directly at the 8-ball (no banking), and succeeds in doing so, then it is almost certain that the cue ball will fall into the other far corner pocket. Suppose, in fact, that under these circumstances it is almost certain that either both balls will fall into pockets or neither will. Keeping the distinction between token and population causation clearly in mind, let us consider the population of attempts in which this novice player shoots the cue ball at the 8-ball without first banking. Let X be the event of the 8-ball dropping into one of the far corner pockets, Y the event of the cue ball dropping into the other far corner pocket, and Z the event of the cue ball colliding with the 8-ball. Suppose also that the probability of the 8-ball's falling into one of the corner pockets, given the player succeeds in striking the 8-ball with the cue ball, is about 0.5, so that the probability of the cue ball's falling into the other corner pocket, given that the player succeeds in striking the 8-ball with the cue ball, is also about 0.5.

The factors Z, X, and Y in this example clearly form an interactive fork: plausibly, they satisfy, (1), (2), (3★), and (4) of the definition, above, of interactive forks. Most pertinently, $Pr(Y/Z\&X) \approx 1 > Pr(Y/Z\&{\sim}X) \approx O$, which is (3★). Also, of course, neither of X and Y is a cause of the other. So we have another example of an interactive fork in which the joint effects are causally neutral for each other.

Let us suppose that, between X and Y, X is the "earlier" factor (that we limit the population to cases in which one of X

and ${\sim}X$ occurs before one of Y and ${\sim}Y$).[30] Then the issue is the probabilistic theory's verdict concerning the causal role of X for Y. Since X is causally neutral for Y, the problem for the probabilistic theory of causation is to find, and justify, a set of background contexts, for evaluating X's causal role for Y, that screen off Y from X. The question is: Can we find factors causally relevant to Y, independently of X, that, when all are held fixed in any given way, screen off Y from X?

Clearly the answer is yes. For one thing, in macroscopic examples such as this one, classical physics assures us that if we describe the collision of the cue ball with the 8-ball in enough detail – specifying the exact relative positions of the balls, the exact direction of motion and momentum of the cue ball, the exact points of contact, and so on – then we can predict with certainty whether or not the balls will fall into the corner pockets. Let Z_i range over these finer descriptions of the collision. Then, holding fixed, positively, any of the Z_i's will confer probabilities of 0 or 1 on X, Y, and $X\&Y$, and Y is screened off from X. This illustrates the fact that we sometimes have to formulate a common cause *partition* of Z_i's in order to recover a conjunctive forklike structure.[31] This also illustrates what Salmon (1984) calls a "perfect fork," a common cause structure in which the probabilities of the joint effects, conditional on the presence or absence of the common cause, are all either 0 or 1. In perfect forks, the common cause always screens its joint effects off from each other, in that there is no correlation between them conditionally on the presence or on the absence of the common cause.

However, in order for the analysis of this example to apply also to examples in which determinism is false, a different

[30] Again, a clarification of the idea of one factor's being earlier in time than another will be given in Chapter 5.

[31] Recall that a *partition* is set of mutually exclusive and collectively exhaustive factors (as noted above and also explained in Appendix 1). The conjunctive fork like structure recovered here is actually a more general kind of conjunctive fork than characterized in (1)–(4) above, where (3) and (4) are replaced by: $Pr(Y/Z_i\&X) = Pr(Y/Z_i\&{\sim}X)$, for each Z_i in the partition.

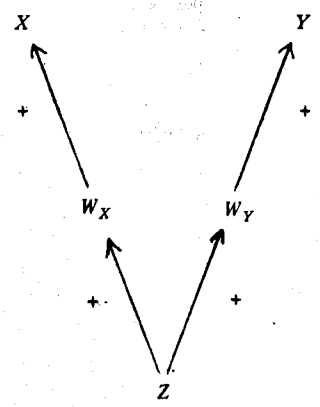


Figure 2.9

kind of approach is needed. Suppose Z occurs, or does not, at a time t_0, and that Y occurs, or does not, at time t_1. Clearly the possible states of the cue ball (including, say, its velocity and direction of motion) *at times between t_0 and t_1* are causally relevant to whether or not Y occurs at t_1 (to whether or not the cue ball falls into a corner pocket at t_1). And these states are causally independent of whether or not X occurs (of whether or not the 8-ball falls into a corner pocket). Let us suppose, for simplicity, that there are just two relevant possible states of the cue ball at a time $t_{1/2}$ between times t_0 and t_1: ball rolling toward the pocket with enough momentum to almost certainly carry it into the pocket (W_Y), and ball rolling in such a way that it is almost certain not to fall into the pocket ($\sim W_Y$).[32] Then Figure 2.9 (ignoring W_X for the moment) de-

[32] A more detailed specification of the intermediate states of the cue ball would specify the possible momenta of the ball, including direction of motion, and so on. In this case, we would have a partition of intermediate factors, or possible intermediate states. Then Z has different kinds of causal significance for the different intermediate factors, and these states have different kinds of causal significances for Y. The simpler analysis given in the text can easily be transformed into such a more detailed analysis.

picts the case in more detail than the first description of it above did.

Since W_Y specifies everything that is, at $t_{1/2}$, causally relevant to whether or not Y will occur at t_1, and since the process from W_Y at $t_{1/2}$ to Y (or to $\sim Y$) at t_1 does not interact with (or "intersect") the process from Z to X (or to $\sim X$) – they are now independent processes – W_Y should screen off Y from X. And the same goes for $\sim W_Y$. And since W_Y is causally relevant to Y, and is causally independent of X, it should be held fixed in background contexts in assessing X's causal role for Y. In this example, an appropriate set of background contexts would be the four ways of holding fixed, positively and negatively, Z and W_Y. And within each of these contexts, Y is probabilistically independent of X; so the probabilistic theory gives the right answer that X is causally neutral for Y.

This example shares a feature noted about the example discussed just before. There is a factor *after* the time of the nonscreening common cause that does screen the joint effects off from each other. However, unlike the previous example, this intermediate causal factor does not, in this example, form a conjunctive fork with the joint effects of the nonscreening common cause. Although this is irrelevant to the adequacy of the theory of probabilistic causation to such cases, it is interesting to note that if we define W_X analogously to the way W_Y was defined (W_X is causally intermediate between Z and X as shown in Figure 2.9), and if we set $W = W_X \& W_Y$, then, plausibly, W, X and Y do form a conjunctive fork.[33] In this case, this example has the structure depicted in Figure 2.8, and is analogous in all formal respects.

Salmon (1978, 1984) also gives a microscopic, and indeterministic, example of an interactive fork. The example involves the role of a conservation law in a phenomenon called

[33] If the correlation between W_X and W_Y is not perfect, then the partition of the four ways of holding these two factors fixed will, with X and Y, form a conjunctive fork, in the more general way of understanding conjunctive forks, noted above.

Compton scattering. Suppose we can consider a given electron to be at rest, and suppose an energetic photon, with energy E, collides with this electron. Call the event of this collision Z. There is a certain probability that a photon will emerge from the collision with energy E_1; call this event X. And there is a certain probability that an electron will emerge from the collision with energy E_2; call this event Y. Suppose E_1 and E_2 add up to E, as the law of conservation of energy demands. Because of this conservation law, there will be a correlation between events X and Y, conditional on Z. For example (as Salmon illustrates the point), if the probability of X given Z is 0.1 and the probability of Y given Z is 0.1, then the probability of $X\&Y$ given Z is not 0.01, the product of the two probabilities, but rather 0.1. This is because the law of conservation of energy (and the fact that E, E_1, and E_2 are related as the law demands) implies that X will occur if and only if Y does.

So $Pr(X\&Y/Z) > Pr(X/Z)Pr(Y/Z)$, which implies (3★) of the characterization of interactive forks above; plausibly (1) and (2) are also satisfied. And Salmon points out that this example, unlike the one involving billiard balls analyzed above, is not susceptible to analysis as a perfect conjunctive fork. The example, Salmon says, is "irreducibly statistical." No more refined or detailed description of the collision will necessitate the emergence of a photon and electron with given energies. However, if this is the *only* relevant difference between the Compton scattering example and the billiard ball example (which it actually is not, as noted below), then the probabilistic theory of causation can avoid the conclusion that X is a cause of Y, or Y a cause of X, in the same way as explained above for the billiard ball example. Whether or not an interactive fork is analyzable as a *perfect* conjunctive fork does not, by itself, control whether or not the probabilistic theory can correctly analyze the causal relations among the components of an interactive fork.

Let us focus on what the probabilistic theory has to say about the causal role of X for Y; exactly parallel considerations will apply to the question of the role of Y for X. Factor Z is a cause of factors X and Y. So, again, in order for the temporal priority requirement (Chapter 5) to be met, Z must be temporally prior to each of X and Y.[34] Let us say again that Z occurs at t_0 and Y occurs at t_1. Z is the event of the collision at t_0 and Y is the event of the electron having energy E_2 at time t_1. Let W_{Yi} range over states of the electron at a time $t_{1/2}$, between times t_0 and t_1. These will specify the energy of the electron at the intermediate time.[35] Each such state is, of course, causally relevant to Y, independently of X; so they must be held fixed in assessing X's causal role for Y. And, of course, conditional on any of these intermediate states, Y is probabilistically independent of X; so again, the probabilistic theory gives the correct answer that X is causally neutral for Y.

Also, if we let W_{Xj}'s range over intermediate states of the photon, then the partition of $W_{Xj}\&W_{Yi}$'s forms a conjunctive fork like structure with X and Y, in the sense that each element of the partition screens X and Y off from each other. Further, a coarser partition of intermediate states would make this example formally equivalent, in terms of factors and probabilities, to the billiard ball example, as depicted in Figures 2.8 and 2.9. The coarser partition would simply disjoin

[34] In discussing this microscopic example involving fast particles, we are coming close to having to take account of relativity of simultaneity and of temporal priority, described in the theory of special relativity. This will be especially pressing when discussing the Einstein–Podolsky–Rosen paradox, below. For now, we may just note that in this example, X and Y each fall in the future light cone of Z, so that each is absolutely future to Z.

[35] So in this example the two particles have definite energies at intermediate times, and I am assuming (falsely) that the example is not of the "Einstein–Podolsky–Rosen" type, which will be discussed below. Whether or not, physically, the Compton scattering phenomenon is of the EPR type is beside the point I want to make here, which is simply that the difference between determinism and indeterminism (or whether or not an interactive fork can be analyzed as a *perfect* conjunctive fork) does not *by itself* control whether or not the probabilistic theory will give the right answers in an interactive fork situation.

those possible intermediate states that are positively causally relevant to each of X and Y into one factor, and disjoin the others into a second factor, the negation of the first.

The billiard ball example and the Compton scattering example are disanalogous in that one is analyzable as a perfect fork and the other is not. However, they are analogous in that they share enough structure for the second analysis given of the billiard ball example to apply also to the Compton scattering example. In each case, the probabilistic theory delivers the correct answers about what causes what.

However, there is another kind of interactive fork possibility for which the analysis is not so clear. These are examples of the "Einstein–Podolsky–Rosen" (EPR) kind. In these examples, roughly, factors Z, X, and Y are described that form an interactive fork, where the joint effects X and Y are spacelike separated, and yet there is no factor (or partition of factors) that describes the state of the system at times before the occurrence of the joint effects and that screens the effects off from each other.[36] This seems especially troublesome for the probabilistic theory of causation, since it seems that we cannot say that either of X or Y is causally relevant to the other, given the "locality" requirement of special relativity theory, understood as meaning that causal processes cannot exceed the speed of light. In the remainder of this section, I will briefly explain some of the issues and some of the bearing of the EPR paradox on the probabilistic theory of causation.

Here is one schematic version of the EPR paradox.[37] Some

[36] Two events are *spacelike separated* if they are outside each other's light cones – that is, if they are so spatially and temporally situated that no subluminal, or luminal, process could originate at the time and place of either of the two events and arrive at the time and place of the other. In this case, neither is absolutely future or absolutely past to the other, according to special relativity theory. If, on the other hand, a subluminal process *could* connect two events, then the two events are said to be *timelike separated,* in which case the two events stand in a definite temporal priority order in all reference frames, according to special relativity theory.

[37] For other discussions from a philosophical point of view, see, for example, the following (on which the discussion of the EPR paradox here is mainly based): Skyrms (1980, 1984b), van Fraassen (1982), Jarrett (1984), Salmon (1984).

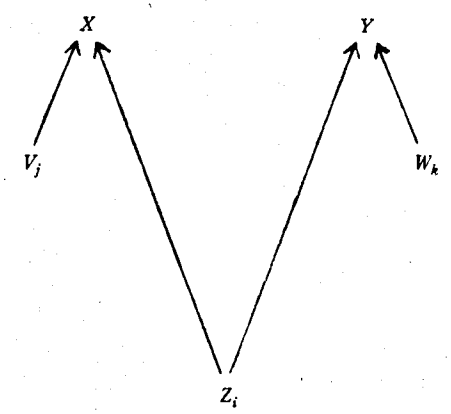


Figure 2.10

event creates two particles that shoot out in opposite directions. At a later time, after the particles are well separated, measurements are made on the two particles. The results of the measurements are called "up" and "down," relative to the spatial orientations of the measuring devices. Let X be that the measuring apparatus on the left gets result "up" ($\sim X$ is "down" on the left), and let Y be that the measuring apparatus on the right gets result "up" at ($\sim Y$ is "down" on the right). Let Z_i range over states of the two-particle system, where the Z_i's can say what the system is like at any time up to, but not including, the time of the two measurements. Finally, let V_j and W_k range over states of the left and right measuring devices, respectively; that is, the V_j's and W_k's range over the different possible spatial orientations of the devices. This much of the causal structure of the example is depicted in Figure 2.10.

Quantum mechanics implies that, given any state of the two-particle system, if the two detectors make measurements in the same direction (that is, with the same spatial orientation), then X and Y will be *perfectly anticorrelated,* and each

will have probability 0.5. For all i and j, $Pr(Y/Z_i\&V_j\&W_j\&X) = 0 < 1 = Pr(Y/Z_i\&V_j\&W_j\&\sim X)$, and $Pr(X) = Pr(Y) = 0.5$. Experimental evidence confirms this. The "paradox" of the EPR paradox is that, according to quantum theory, there is this anticorrelation between factors X and Y (or correlation between $\sim X$ and Y), where it seems that neither factor can be causally relevant to the other (since spacelike separated), and where the correlation does not disappear when all factors causally relevant to the events X and Y are specified. Bell (1964, 1971) showed that no theory – no matter what causally relevant factors (or "hidden variables") are admitted into the states Z_i of the system (as long as the hidden variables do not depend on the settings of the detectors) – could *both* give independence of the two measurement results, conditional on the state of the system and detector settings, *and* predict the actual experimental and quantum mechanical statistics. Let me now briefly clarify these ideas.

One kind of "locality" condition, inspired by special relativity theory, is that the result on the right-hand measurement device should be probabilistically independent of both the setting and the result on the left, conditional on a state of the two-particle system and on the state of the measurement device on the right – and the same in the other direction, from right to left. The state of the two-particle system, together with the orientation of the right-hand measurement device, should screen off the result on the right both from the setting on the left and from result on the left, and the same in the other direction, according to this understanding of locality. Roughly following Jon Jarrett (1984), let us call this the condition of *strong locality* (the notation here differs from Jarrett's):

For all i, j, and k,

$$\begin{aligned} Pr(Y/Z_i\&W_k) &= Pr(Y/Z_i\&W_k\&V_j\&X) \\ &= Pr(Y/Z_i\&W_k\&V_j\&\sim X) \\ &= Pr(Y/Z_i\&W_k\&V_j) \\ &= Pr(Y/Z_i\&W_k\&X) \end{aligned}$$

$$\begin{aligned} &= Pr(Y/Z_i\&W_k\&\sim X), \text{ and} \\ Pr(X/Z_i\&V_j) &= Pr(X/Z_i\&V_j\&W_k\&Y) \\ &= Pr(X/Z_i\&V_j\&W_k\&\sim Y) \\ &= Pr(X/Z_i\&V_j\&W_k) \\ &= Pr(X/Z_i\&V_j\&Y) \\ &= Pr(X/Z_i\&V_j\&\sim Y). \end{aligned}$$

(Of course, there is some redundancy, on the right-hand sides, in this formulation of strong locality; for example, the first, second, sixth, and seventh lines alone would be sufficient to characterize the idea.) Quantum mechanical predictions violate strong locality. Einstein, Podolsky, and Rosen (1935) concluded that quantum theory must be *incomplete* – that there must be factors that the theory has failed to take account of ("hidden variables") that, if included in the Z_i's, would yield the independence embodied in the strong locality condition.

However, experimental data is in harmony with the predictions of quantum mechanics. And Bell's theorem (1964, 1971), mentioned above, implies that, as long as any new hidden variables do not depend on the V_j's and W_k's, there can be no probability function describing the experiment described above that *both* satisfies strong locality *and* predicts the quantum mechanical and experimental statistics. Given special relativity, it is very plausible that the hidden variables, included in the Z_i's, should not depend on the settings of the measuring devices, the V_j's and W_k's. This is because the devices can be set, each in a *random manner,* and each at "the last moment," at space-time points that are *spacelike separated,* both from each other and from the particles. As Skyrms (1984b) points out, the choices of settings can be made by separate indeterministic quantum mechanical devices whose operations we have every theoretical reason to believe *are* independent of the state of the two particle system.

Strictly speaking, however, what special relativity precludes is superluminal transmission of *information-carrying signals* – that is, the transfer of such signals between spacelike

separated space-times. The idea of a "signal" has what has been called a "broad" and a "narrow" connotation. In the broad sense, a signal can be transmitted if there is some kind of *correlation* between the *outcome* on one apparatus and the *outcome* on another. In the narrow sense, a signal is transmitted if there is a *transfer of information* from one apparatus to another. This distinction between the broad and narrow senses of signal has a precise formulation, which I will explain below. Special relativity only precludes superluminal signals in the narrow, information-carrying sense of signal. Because of this, quantum mechanic's violation of strong locality is compatible with special relativity, as I shall now explain.

Suppose two experimenters are stationed at the spacelike separated space-times of the left- and right-hand measurements in the experiment described above, one experimenter on the left and the other on the right. The two experimenters can decide the *orientations* of their respective detectors, but (according to quantum theory) they cannot control the *outcomes* on their detectors. If, given a state Z_i of the two particle system and a state W_k of the right-hand detector, there were a correlation between the *state,* V_j, of the left-hand detector and the *outcome,* Y or $\sim Y$, on the right-hand side, then the experimenter on the left could transmit (at least in a probabilistically reliable way) a superluminal signal about the state of his detector to the experimenter on the right. (The state Z_i of the two-particle system and the state W_k of the right-hand detector could, theoretically, be prearranged.) This would be an information-carrying signal, and thus a signal in the narrow sense of signal. Special relativity rules out this kind of signal.

In addition to ruling out this kind of correlation, between *state* on one side and *outcome* on the other (conditional on any state of the two-particle system and any state of the detector on the other side), strong locality also rules out correlation between *outcome* on one side and *outcome* on the other (conditional on any state of the system and any state of the detector on the other side). While the first kind of correlation is ruled

out by special relativity, the second is not. This is because although an experimenter on one side can decide the *state* of his detector, he cannot control the measurement *outcome,* and thus *cannot use the outcome to send a signal to the other experimenter,* in the "narrow" sense of "signal" that means "transfer of information."[38]

In light of special relativity, this suggests a locality requirement that is weaker than strong locality, and, in light of the success of quantum mechanics, more plausible: Given any state Z_i of the two-particle system and any setting W_k of the right-hand measuring device, the measurement result on the right, Y or $\sim Y$, is independent of the *setting* V_j on the left-hand device – and the same in the other direction, from right to left. This means that, conditional on any state of the two-particle system and any state of one of the detectors, the *result* of the measurement on the one detector is independent of the *state* of the other detector, but not necessarily of the *result* of the other measurement.

Again roughly following Jarrett (1984), let us call this the condition of *weak locality:*

> For all i, j, and k,
> $Pr(Y/Z_i\&W_k) = Pr(Y/Z_i\&W_k\&V_j)$, and
> $Pr(X/Z_i\&V_j) = Pr(X/Z_i\&V_j\&W_k)$.

The difference between strong and weak locality is that strong locality requires independence of the outcome on one side from *both settings and outcomes* on the other side, while weak locality only requires independence of the result on one side from the *setting* on the other – all conditional on a state of the two-particle system and a setting on the one side.

The predictions of quantum mechanics are in harmony not only with experiment, but also with weak locality. But as mentioned above, these predictions violate strong locality. Jarrett (1984) shows that strong locality is logically equivalent

[38] Jarrett (1984) reports a proof by Ghirardi, Rimini, and Weber (1980) that no superluminal signal can be produced by quantum mechanical measurements.

to the conjunction of weak locality and a condition he calls "completeness":

For all i, j, and k,

$$\begin{aligned} Pr(Y/Z_i\&W_k\&V_j) &= Pr(Y/Z_i\&W_k\&V_j\&X) \\ &= Pr(Y/Z_i\&W_k\&V_j\&\sim X), \text{and} \\ Pr(X/Z_i\&V_j\&W_k) &= Pr(X/Z_i\&V_j\&W_k\&Y) \\ &= Pr(X/Z_i\&V_j\&W_k\&\sim Y). \end{aligned}$$

(Of course, there is some redundancy, on the right-hand sides, in this formulation of completeness; for example, the first and third lines alone would suffice to characterize the idea.[39]) Quantum mechanical predictions violate completeness. If we hold fixed the state of the two-particle system, and set both detectors in the same direction, then X and Y are perfectly anticorrelated, though this correlation does not make signaling possible.

What is the bearing of this on the probabilistic theory of causation? If we consider a population of experiments in which the state of the two-particle system does not vary, and the orientations of the two detectors are the same and do not vary, then the probabilistic theory tells us that X is a negative causal factor for Y in this population, since in this population $Pr(Y/X) = 0 < 1 = Pr(Y/\sim X)$ (note that this population is homogeneous with respect to all factors causally relevant to Y independently of X).[40] There are several possibilities open

[39] It is easy to see that strong locality is equivalent to the conjunction of weak locality and completeness, given the way in which strong locality and completeness have been redundantly formulated here. *Suppose strong locality.* Weak locality is just the equalities, for i, j, and k, between the first and fourth probabilities and between the fifth and eighth probabilities given in the formulation of strong locality above, which equalities follow by transitivity of equality. And completeness is just the equalities, for all i, j, and k, among the probabilities in the first triple, and among the probabilities in the second triple, of right hand sides in the formulation of strong locality above. *Now suppose weak locality and completeness.* Let i, j, and k be arbitrary, for the proof of strong locality. $Pr(Y/Z_i\&W_k) = Pr(Y/Z_i\&W_k\&V_j)$, by weak locality; and $Pr(Y/Z_i\&W_k\&V_j) = Pr(Y/Z_i\&W_k\&V_j\&X)$, by completeness. This gives us that $Pr(Y/Z_i\&W_k) = Pr(Y/Z_i\&W_k\&V_j\&X)$, the first line of the formulation of strong locality above. The rest of this part of the proof follows the same pattern.

[40] It is perhaps worth noting that, given quantum mechanical statistics, when the independent causally relevant factors (the Z_i's, the V_j's, and the W_k's) are allowed to vary (specifically, when the orientations of the detectors with respect to each other differ from each other in different ways), then X's probabilistic (or "nonlocal causal") role for Y will vary as well.

as to how assess, or adjust, the probabilistic theory in light of this.

If we wish to deny that X is causally relevant to Y in the example, then we could include in the theory, by fiat, the condition that one event can only be causally relevant to events that lie in the first event's future light cone, a possibility noted by Skyrms (1984b). However, it seems that this approach cannot really get to the heart of the matter, that it cannot make contact with the reasons why one would wish to deny that X is causally relevant to Y in the example. When X occurs, so does $\sim Y$. Now suppose $\sim Y$ is, in a noncontroversial way, positively causally relevant to a third event, U, that *is* within the future light cone of X; and suppose that X is otherwise irrelevant to U (for example, Y and $\sim Y$ each screen off U from X). The requirement that effects must lie within the future light cones of causes will not prevent us from saying that X is a cause of U. But it seems that the idea that X causes U in this example should be just as unsatisfactory as the idea that X is causally relevant to Y, for those who wish to deny that X is causally relevant to Y.

Another possibility is to deny that causation must be local – in the strong sense of locality, of course, whose denial is consistent with weak locality, does not imply the possibility of superluminal, information-carrying signals, and is thus not in conflict with special relativity. In this case, the probabilistic theory would give what may be the *right answer,* namely, that X *is* causally relevant to Y: In the population described above, X *is* causally negative for Y. Note that if we adopt this position, then we are forced to say also that Y is causally negative for X, violating asymmetry of causation. This is because (1) the population is homogeneous with respect to all factors that have to be held fixed in assessing the causal role of either of X or Y for the other, (2) correlation is symmetric,

and (3) neither event is (in special relativity theory) "absolutely" temporally prior to the other, the two events being spacelike separated. This third fact means that, if we agree that X is causally relevant to Y, then we cannot rule out the idea that Y is causally relevant to X on the basis of considerations involving the temporal order of events. This follows unless we say (implausibly, it would seem) that causal relevance is relative to frame of reference.

A denial of strong locality requires a little care in the formulation of a temporal priority requirement. In Chapter 5, the requirement will read roughly like this: *If* X and Y are *timelike* separated (so that they are within each other's light cones), *then* X can be causally relevant to Y only if X is before (and thus absolutely before) Y. This formulation does not rule out symmetry of causation between spacelike separated events, but it does rule out "absolutely backwards" causation.[41] Of course, a denial of (strong) locality does not come with an account of a *mechanism* of nonlocal causation (and the idea of physical mechanisms does not explicitly enter into the theory of type level probabilistic causation anyway). But this would nevertheless seem to be one consistent and coherent way of developing one concept of cause.[42]

Clearly, all this cannot be settled here, since the question of the possibility of *physical mechanisms* behind nonlocal "connections" is so highly relevant, and since only physical theory can address this question.[43] I have nothing more to say about the bearing of EPR phenomena on the probabilistic theory of causation – except that it also seems clear that it would be premature to conclude that the probabilistic theory of causation must exclude EPR phenomena, or to conclude that it simply does not apply to these cases.

[41] It is perhaps worth reiterating here that the delicate idea of ordering event *types* in time will be clarified in one way in Chapter 5.
[42] Compare Skyrms (1984b).
[43] Compare Salmon (1984, pp. 258–9).

3

Causal interaction and probability increase

For the examples of spurious correlation discussed in Chapter 2, it sufficed to hold fixed all (independent) positive, negative, and mixed *causes* of the candidate effect factor, in order for the probability-increase idea to deliver the right answers about what caused what. For these examples, only factors that were *causally relevant* to the candidate effect factor needed to be held fixed. In this chapter, I will argue that other kinds of factors, which may be causally *irrelevant* to (*neutral* for) the effect factor in question, must be held fixed as well, if the probability-increase theory is to deliver the right answers in other kinds of cases.

For example, if the right answer in Dupré's example, discussed in Chapter 2, is that smoking has a *mixed* causal role (not positive, negative, or neutral) for lung cancer, then it will be necessary to hold fixed the factor of that rare physiological condition. Otherwise, causal relevance would go by *average* probabilistic impact of smoking on lung cancer, across the presence and absence of that condition, and this cannot give the correct answer of mixed causal relevance. However, as noted in Chapter 2 and explained more fully in this chapter, that physiological condition need not itself be a positive, negative, or mixed cause of lung cancer.

At the beginning of Chapter 2, the possibility of there being such factors as that physiological condition in Dupré's example was called *the problem of causal interaction*. In Section 3.1, I give a simple formulation of the idea of interaction that characterizes such cases, and I argue that interacting causal factors must be held fixed in assessing causal roles.