1

Extrapolation and Heterogeneity

Genuine philosophical problems are always rooted in urgent problems outside philosophy, and they die if these roots decay.

—Karl Popper[1]

The best way to introduce the topic of this book is with a few examples.

- Studies find that a particular substance is a carcinogen in rats. We would like to know whether it is also such in humans.[2]
- A randomized controlled experiment has found that a pilot welfare-to-work program improved the economic prospects of welfare recipients. It is desired to know whether the program will be similarly effective in other locations and when implemented on a larger scale.[3]
- On the basis of a controlled experiment concerning outcomes resulting from initiating anti-retroviral therapies earlier or later among HIV+ patients, a physician wishes to decide the best time to initiate this therapy for the patients she treats.

In each of these cases, one begins with some knowledge of a causal relationship in one population, and endeavors to reliably draw a conclusion concerning that relationship in a distinct population. I will use the term *extrapolation* to refer to inferences of this sort. If the populations in question were perfectly homogeneous, extrapolation would be easy: the result from the first population could be directly carried over to the second. But it is not reasonable to assume that the populations in the foregoing examples are homogeneous: they almost certainly differ with respect to characteristics that affect the causal relationship in question. I will use the expression *the problem of extrapolation in heterogeneous populations* (or *the problem of extrapolation* for short) to refer to the challenge of transferring causal generalizations from one context to another when homogeneity cannot be presumed.

The motivation for extrapolation is that evidence concerning the model, or base, population is often more accessible than that for the target with which one is chiefly concerned. For instance, there are many experiments that can be performed on animal models that cannot, for obvious ethical reasons, be performed on humans. However, this only provides a reason why one would want to extrapolate, and does not explain how experimental results concerning a model can be legitimately transferred to a target. Causal relationships in biology and social science typically depend on a variety of conditions that are subject to change, and

it is rare that all such factors are known and can be measured. As a consequence, a causal generalization that holds in a given population may be false of a subpopulation or other related populations. The effectiveness of a welfare-to-work program depends on an array of features of the individuals involved, as well as the economic conditions of their locality. Likewise, the effect of an anti-retroviral HIV therapy depends on a range of features of the host as well as of the details of the strain, or strains, responsible for the infection. In both cases, it is unlikely that all of the factors upon which the causal relationship depends can be taken into account in an analysis of the problem. Moreover, the examples provided above illustrate the relevance of extrapolation to such policy issues as regulating the use of a chemical or reforming a social program. This book explores how and under what circumstances reliable extrapolation is possible in biology and social science, and explores some of the implications of this topic for issues in philosophy of biology and social science.

There are several strategies that one might take with regard to extrapolation. The most straightforward is what can be called *simple induction*: infer that a causal relationship found in one population holds approximately in other related populations unless there is some reason to suppose otherwise. Although simple induction is an undeniably important aspect of extrapolation, its limitations are well documented in the toxicology literature (Gold et al. 1992; Hengstler et al. 1999). Simple induction often yields mistaken extrapolations, and it provides no guidance when there is some reason to suspect that the extrapolation might be incorrect. The question is whether there is a more sophisticated account of extrapolation capable of overcoming these limitations.

There are two basic challenges that confront any account of extrapolation that seeks to resolve the shortcomings of simple induction. One challenge, which I call *extrapolator's circle*, arises from the fact that extrapolation is worthwhile only when there are important limitations on what one can learn about the target by studying it directly. The challenge, then, is to explain how the suitability of the model as a basis for extrapolation can be established given only limited, partial information about the target. Critics of animal extrapolation sometimes present this challenge in the form of a vicious circle: establishing the suitability of the model would require already possessing detailed knowledge of the causal relationship in the target, in which case extrapolation would be unnecessary. The second challenge is a direct consequence of the heterogeneity of populations studied in biology and social science. Because of this heterogeneity, it is inevitable that there will be causally relevant differences between the model and the target population. Thus, an adequate account of extrapolation must explain how it can be possible to extrapolate from model to target even when some causally relevant differences are present. Both of these challenges have been posed as general critiques of the methodology of animal extrapolation (cf. LaFollette and Shanks 1993a, 1993b, 1995,

I argue that earlier work has answered neither of these challenges. There is a small literature that discusses methodological issues relating to extrapolation by reference to detailed case studies (cf. Burian 1993; Ankeny 2001; Schaffner 2001; Guala 2005; Alexandrova 2006). These authors point out that extrapolation is on firmer ground with regard to highly conserved mechanisms, and that the suitability of a model for a particular extrapolation is an empirical hypothesis that must be supported by evidence. But although such methodological observations are undoubtedly correct, they answer neither of the challenges described above. Difficult cases of animal extrapolation typically concern causal relationships—such as the carcinogenic effect of a particular compound—that are not highly conserved. And the observation that the suitability of a model for extrapolation is an empirical hypothesis does not answer the extrapolator's circle. How can that empirical hypothesis be established without already knowing what one wanted to extrapolate? Nor do these studies indicate how extrapolation can be justified when there are some causally relevant differences between model and target.

Others have proposed that *capacities* or *causal powers*—understood as stable influences that are relatively independent of context—can serve as a basis for extrapolation (Cartwright 1989, 1999; Cheng 1997, 2000). The difficulty here is that questions of the stability versus context dependence of a causal relationship are precisely what are at issue in cases of extrapolation. I argue (in Chapter 5) that, when pressed on this matter, existing proposals concerning capacities and causal powers either to revert to simple induction or morph into a version of the mechanisms approach. The mechanisms approach rests on the intuition that knowing how a cause produces its effect provides can provide a basis for extrapolation. It proposes that knowledge of the mechanisms running from cause to effect and of the kinds of things that can interfere with them enhances our ability to reliably decide whether a causal relationship found in one population will or will not obtain in another. This thought is second nature among molecular biologists, and several authors concerned with philosophical questions regarding the role of mechanisms in science have suggested it in passing (cf. Wimsatt 1976, 691; Stinchcombe 1991, 367; Elster 1998, 49; Schelling 1998, 36–37). Yet without further elaboration, the mechanisms proposal does not answer the two challenges described above either. It does not answer the extrapolator's circle, since it is unclear how one can show that the mechanisms in the model are similar enough to the target to justify extrapolation, given the limitations on one's ability to study the mechanisms in the target directly. Moreover, some differences in the mechanism in the model and target are inevitable in biological and social science examples. Thus, the mere invocation of mechanisms does not explain how extrapolation can be justified in the presence of causally relevant disanalogies between model and target.

In this book, I further develop the mechanisms approach to extrapolation so as to more adequately respond to these challenges. I endeavor to

larify the premises that underlie applications of the approach and the ypes of extrapolative inferences these premises can support, as well as he relevance of extrapolation to some familiar topics in the philosophy of iology and social science. I believe that this project is valuable for several easons. First, the project is of practical relevance to scientific method- logy. A clear understanding of the premises underlying the mechanisms pproach to extrapolation helps to reveal the possibilities and limitations f this strategy. If there are circumstances when the requisite premises are roblematic, then it is important to know this so as to avoid unreliable pplications of the approach. On the other hand, there may be inferences hat would be justified by the mechanisms approach that are not being aken advantage of, and an examination of basic assumptions may show in what ways this is the case. Second, I believe that the project is of significant interest for more traditional philosophical topics. As I endeavor to show, a variety of familiar philosophical issues in biology and social science are linked to the problem of extrapolation in heterogeneous populations, including reductionism, ceteris paribus laws, and causality.

The organization of the book is as follows. Chapter 2 presents and explicates a set of concepts—intervention, causal effect, and causal relevance—that recur throughout the remainder of the book. Although my analysis of the first two of these concepts is mostly drawn from other authors, my discussion of causal relevance makes an original contribution insofar as proposing a definition of positive and negative causal relevance that is applicable to cases in which the cause and effect are represented by quantitative variables. It is sometimes claimed that a definition of positive causal relevance should include a criterion of *contextual unanimity*, which requires that a positive causal factor raise the probability of the effect in all background contexts. I argue that this is a mistake, and that such criteria should not be regarded as part of the *meaning* of causal relevance but rather as circumstances that may, when present, facilitate extrapolation.

A mechanisms approach to extrapolation requires an account of the relationship between the qualitative concept of a mechanism and the probabilistic causal notions described in Chapter 2. Chapters 3 and 4 address this issue. The main proposal of Chapter 3 is an account of how mechanisms, on the basis of domain-specific arguments, can be identified with *causal structure*. Often represented by directed graphs, causal structure is that which generates probability distributions and provides information concerning how these distributions will change under interventions. An argument for identifying mechanisms with causal structure in a given context takes the form of an *empirical analysis*, a concept that I draw, with some modifications, from Phil Dowe (2000). A central part of this empirical analysis consists of providing reasons to think that mechanisms are modular in the sense of having independently changeable components. I explain how evolutionary theory can be used to motivate the premise that mechanisms in molecular biology are modular.

natural selection may favor modularity, very little has been written on this topic in social science. I make some suggestions about how this evolutionary argument might transfer to social science, but conclude that the case for identifying mechanisms with causal structure is, at present, less well founded in social science than in molecular biology. Further discussion of the circumstances under which social mechanisms can be identified with causal structure is deferred until Chapter 8.

Although the identification of mechanisms and causal structure is an important element of mechanisms-based extrapolation, it is only a first step. Many, and perhaps most, applications of the approach require further, more specific premises about the relationship between probability and causal structure, and hence mechanisms. Chapter 4 articulates a proposition, labeled the *disruption principle*, which plays a fundamental role in mechanisms-based extrapolation of probabilistic causal claims. The disruption principle asserts that interventions on a cause make a difference to the probability of the effect if and only if there is an undisrupted mechanism running from the cause to the effect. After presenting the disruption principle in the abstract, I illustrate it by means of an example drawn from HIV research. Next I consider what justification can be given for the disruption principle. I show that, given the identification of mechanisms with causal structure, it can be derived from two more familiar principles concerning probability and causality, namely, the *principle of the common cause* (PCC) and the *faithfulness condition* (FC). I argue that the aspect of the PCC relevant to the disruption principle rests upon very solid ground but that the case of the FC is more complex. A common objection to the FC is that it is likely to be false when there are counteracting causal paths. I show that such arguments are valid only given a further condition that rarely obtains in heterogeneous populations. Nevertheless, there are some circumstances—such as gene knockout experiments—in which exceptions, or at least *near* exceptions, to the FC are a more serious concern. Hence, this discussion identifies a potential limitation of the disruption principle, and thereby of the mechanisms approach to extrapolation.

Chapters 5 and 6 utilize the concepts articulated in the foregoing chapters to develop an account of mechanisms-based extrapolation. Chapter 5 begins by examining the limitations of extrapolation by simple induction. Next I argue that previously proposed versions of the capacities and mechanisms approaches do not adequately address the two challenges mentioned above: the extrapolator's circle and explaining how extrapolation can be justified in the presence of causally relevant differences between model and target. I then proceed to develop the mechanisms-based answers to these challenges. I begin by explaining how a mechanism in a model organism might serve as a basis for inferring the existence of a corresponding mechanism in the target, by means of what I call *comparative process tracing*. Comparative process tracing relies on background knowledge concerning stages of the mechanism at which

significant differences are likely to occur, and where such differences are not likely. In addition, the number of points that must be compared can be reduced further by focusing on downstream stages of the mechanism. Comparative process tracing answers the extrapolator's circle, then, by showing how limited knowledge of the mechanism in the target can suffice to establish the suitability of the model as a basis for extrapolation. I illustrate comparative process tracing by reference to the case of aflatoxin B_1, which concerned the extrapolation of a carcinogenic effect from rodents to humans. Finally, I briefly discuss the relevance of my proposal for disputes concerning the ethical justifiability of animal research.

The second basic challenge confronting an account of extrapolation in heterogeneous populations is that it must explain how extrapolation can be possible even when there are causally relevant differences between model and target. My answer to this challenge is first proposed in Chapter 5, in connection with the aflatoxin B_1 example, and is developed in further detail in Chapter 6, in the context of a discussion of ceteris paribus laws. The central point is that the closeness of match required between model and target depends upon the specificity of the causal claim that one wishes to extrapolate. In particular, a total absence of causally relevant disanalogies is *not* required for extrapolating claims about positive and negative causal relevance. That point is illustrated by the aflatoxin example in Chapter 5, and Chapter 6 articulates some sufficient conditions for extrapolating positive or negative causal relevance. Chapter 6 also discusses a philosophical issue that is closely associated with extrapolation, namely, ceteris paribus laws, which are laws qualified by a clause to the effect of "other things being equal" or "so long as nothing interferes." The expression "ceteris paribus law" is in fact highly ambiguous. Some types of generalizations labeled "ceteris paribus laws" are unproblematic, while the opposite is true for one common interpretation. I explain how the infirmities of the most problematic type of ceteris paribus law vanish if "ceteris paribus" is interpreted as qualifying the extrapolation of positive causal relevance rather than the truth of a universal generalization.

The mechanisms approach to extrapolation is also linked to a perennial issue in philosophy of biology, namely, reductionism. The mechanisms approach to extrapolation operates on the implicit assumption that lower-level details are the place to look for explanations of exceptions to higher-level generalizations. That perspective seems closely tied to reductionism, yet that connection is potentially worrisome, given that reductionism is a highly contentious doctrine. In Chapter 7, I explicate the link between mechanisms-based extrapolation and reductionism. I begin with the suggestion that there are several possible motives for reduction and that different versions of reductionism can be distinguished on the grounds of which goals they aim to achieve. This discussion is used as a basis for clarifying which types of reductionism are presently defended. I then propose that mechanisms-based extrapolation

proposition I call *corrective asymmetry*. Corrective asymmetry obtains when one level of description plays a special role in correcting generalizations at another level, a corrective role which is not reciprocated. I maintain that corrective asymmetry is a criterion of what makes one level more fundamental than another, and hence is a basis for identifying which forms of reductionism genuinely deserve the name. But I also argue that some forms of reductionism that entail corrective asymmetry are compatible with pluralism. In fact, I suggest that corrective asymmetry is helpful for explicating the pluralistic idea that there are autonomous levels of explanation.

Since the best examples of the mechanisms approach to extrapolation I know of come from the biological sciences, the account of extrapolation I propose is developed first in relation to case studies drawn from that domain. Chapters 8 and 9 take up the question of whether the mechanisms approach to extrapolation can be fruitfully extended to social science. There are several challenges confronting this methodological transfer. One is the possibility that social mechanisms do not satisfy the conditions required of causal structure. Chapter 8 picks up the thread of this discussion from Chapter 3. I articulate the concept of structure-altering intervention and explore the circumstances in which interventions are most likely to produce nonmodular changes in social mechanisms. I then turn to a social science example in which extrapolation was a serious concern, namely, the attempt from the mid-1980s to 1990s to estimate the effect of broad-scale changes to the U.S. welfare system on the basis of demonstration programs. Owing to its large scale and relatively unprecedented nature, the intervention in this case was likely to be structure-altering. And in fact, there was a methodological dispute surrounding these studies concerning the value of mechanisms for extrapolating results. I show that a thoroughgoing mechanisms approach, as described in Chapters 5 and 6, is unlikely to be applicable in this case. Nevertheless, I suggest that examinations of social mechanisms in the welfare example are an important supplement to simple induction.

A second challenge for the mechanisms approach to extrapolation in social science is uncertainty about what mechanisms are present. This point is illustrated in Chapter 8 by a case study drawn from experimental economics. The case concerns the extrapolation of a phenomenon known as "preference reversal" from the laboratory to real-world contexts. I show how which of two possible mechanisms is correct has significant implications for how widespread preference reversals are outside the laboratory walls. Chapter 9 examines the challenge of reliably learning social mechanisms. Several authors (cf. Darden and Craver 2001, 2002; George and Bennett 2005) have advanced *process tracing* as a means for discovering mechanisms in biology and social science. Several authors have claimed that process tracing is distinct from and supplements causal inference from statistical data. I argue that existing accounts of how

process tracing overcomes challenges confronting causal inference from statistical data in social science are unsuccessful. I then propose a more adequate account that is based on the insight that the appropriate contrast with process tracing is not causal inference from statistical data but rather what I call *direct causal inference*.

2

Interventions, Causal Effects, and Causal Relevance

This chapter presents several concepts—namely, those listed in the chapter title—concerning causality and probability that play a fundamental role in the treatment of extrapolation in heterogeneous populations developed in the remainder of the book. The concept of an intervention has been discussed at length by other authors (cf. Woodward 1999, 2000, 2003; Hausman and Woodward 1999; Spirtes, Glymour, and Scheines 2000), and my presentation of the topic mostly follows these sources. Likewise, I use Judea Pearl's (2000) definition of causal effect, according to which a causal effect of X upon Y in a population P is a function specifying the conditional probability distribution in P of Y, given interventions that set X to specific values.

My development of the concepts of positive and negative causal relevance, in contrast, is an original contribution. One important type of extrapolation problem has the following form: We know that X is a positive causal factor for Y in the population P, and we want to know whether it is also such in the distinct population P′. A systematic inquiry into this inference problem requires a precise and general definition of the expression "positive causal factor." However, such a definition is not to be found in the literature. Philosophical examinations of causal relevance typically treat causality as a relation between events that occur or do not occur, or between properties that are present or absent. Yet many causal relationships of interest to science and ordinary life hold among factors that are naturally represented as varying on a numerical scale: interest rates and rate of inflation; years of education and income; LDL cholesterol level and arterial constriction; fertilizer dosage and plant growth; and so on. Variables representing features of this sort may be called *quantitative*, in contradistinction to those that merely indicate the presence or absence of a property or occurrence or nonoccurrence of an event, which may be called *qualitative*. Christopher Hitchcock (1993, 1995) has shown, though without quite putting it this way, that some traditional philosophical puzzles concerning causal relevance arise from attempting to characterize causal relationships among quantitative variables by means of definitions of causal relevance that are appropriate only for qualitative variables. Hitchcock proposes that claims concerning causal relevance should be understood as providing qualitative information about the causal effect in question (1993, 350).

Although I generally agree with Hitchcock's proposal as far as it goes, t leaves unanswered several questions that need to be resolved before my pproach to the problem of extrapolation can proceed. In the case of quantitative variables, just what information concerning the causal effect s provided by expressions that indicate positive (or negative) causal elevance? Moreover, how does an account of positive and negative causal relevance for quantitative variables connect to that for qualitative variables? Presumably, the definition for qualitative variables should be a special case of the one for quantitative variables, but just how is that to work? I undertake to develop an account of causal relevance that answers the above questions. Finally, I consider the suggestion that a requirement known as *contextual unanimity* should be added to any definition of positive and negative causal relevance. I argue that such an amendment would be inappropriate.

2.1 INTERVENTIONS

Interventions are manipulations of something, typically with the intention of bringing about further changes in something else. An intervention might be a complex surgical procedure, the simple act of flipping a switch, or the Federal Reserve's decision to cut interest rates by a quarter of a percent. The concept of an intervention is very useful for drawing the distinction between causation and correlation, a point which can be illustrated by means of an old and familiar example.

We know that there is a statistical association between barometric readings and the occurrence of storms. Let B, A, and S be variables representing barometer readings, atmospheric pressure, and the occurrence of storms, respectively, and let the arrow represent the relationship of direct causation. Of course, the notion of direct causation is relative to the set of variables under consideration, since intermediate causal nodes could be added indefinitely through a continually finer-grained analysis. Then we think that the association between B and S is due to their being effects of the common cause A. Directed graphs consisting of nodes linked by arrows, as in Figure 2.1, will be used throughout to depict *causal structures*. In a directed graph, nodes represent variables (e.g., barometer reading, atmospheric pressure, etc.), while the arrows represent the relationship of direct causation and the absence of an arrow indicates the absence of any causal influence. Thus, the graph in Figure 2.1 says that atmospheric pressure is a direct cause of both barometer readings and storms, but that barometer readings have no influence on storms nor storms upon barometer readings. Causal structures and their relationship to mechanisms will be the topic of discussion in Chapter 3. For the moment, however, a rough characterization of causal structures will have to do. Causal structures refer to complexes of cause-and-effect relationships, as embodied in such things as the electrical wiring in a

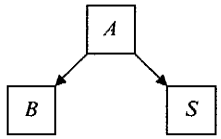


Figure 2.1 Correlation due to a common cause

like that in Figure 2.1 claims to accurately represent some aspect of a causal structure, in this case, one involving a barometer and a meteorological system.[1]

Although the barometer reading is correlated with the occurrence of storms, making it possible to use B to imperfectly predict S, we do not think that B causes S. Part of what this judgment means is that the association between B and S would disappear if we were to intervene as follows. Suppose we place the barometer in a chamber whose air pressure can be set at will, thus allowing us to fix the barometer's reading at any desired level completely independently of A. For example, we could randomly choose numbers in the range of possible barometric readings, and then set the reading of the barometer at these values through manipulations of the pressure within the chamber. Under these circumstances, we would expect the probabilistic dependence between B and S to vanish. On the other hand, if the state of the barometer were (strangely enough) a cause of storms, then we would expect that we could alter the chance of storms by manipulating B. This is a commonsense insight regarding causality: interventions on causes yield changes in effects, but not vice versa.

The general concept of an ideal intervention can be abstracted from this simple example. One begins by finding a source of exogenous variation, such as a purely random process such as a coin toss or a roll of dice. The source of variation is exogenous in the sense that, except under the special conditions implemented in the experiment, it is entirely unrelated to the causal process being studied. It comes, as it were, from the "outside." For example, under normal circumstances, barometer readings, atmospheric pressure, and storms are all completely independent of the outcomes of coin flips or rolls of dice. The intervention then consists of arranging the situation so that the source of exogenous variation determines the value of one of the variables in question. For example, given the intervention described in the preceding paragraph, the barometer is no longer affected by the atmospheric pressure, but only by the randomly assigned air pressure inside the vacuum chamber.

An *ideal intervention* can be defined in the following way. Let $\mathbf{V}$ be a set of variables relevant to a causal structure of interest. Then:

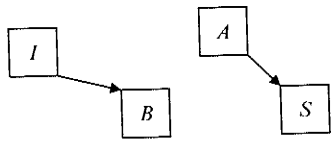
> *Definition 2.1 (Ideal Intervention):*[2] I is an ideal intervention on $X \in \mathbf{V}$ if and only if it is a direct cause of X that satisfies these three conditions:

(a) I eliminates other influences upon X but otherwise does not alter the causal relations among $\mathbf{V}$.
(b) I is a direct cause of no variable in $\mathbf{V}$ other than X.
(c) I is exogenous.

The intervention is exogenous with respect to $\mathbf{V}$ just in case it is neither an effect of any variable in $\mathbf{V}$ nor shares a common cause with any variable in $\mathbf{V}$. Intuitively, exogenous causes come from "outside" the system. As the barometer example illustrates, randomization is a common way of ensuring that the intervention is exogenous. The intervention in the barometer-storm example can be represented graphically, as in Figure 2.2.

The graph in Figure 2.1 can be called the "pre-manipulation graph," and that in Figure 2.2, the "post-manipulation graph." Note that all three requirements of the definition of an ideal intervention are satisfied in this case. First, the intervention fully determines the value of B, removing all other causal influences, which is represented in this case by the deletion of the arrow from A to B. But aside from obviating any other influences upon the target variable (in this case, B), the intervention leaves all other causal relationships in the original graph unchanged. For example, in Figure 2.2, an arrow from A to S is present, just as in Figure 2.1. Second, the intervention is not a direct cause of any member of the set $\{B, A, S\}$ besides B. Finally, the intervention is exogenous, since it is not an effect of any of these variables nor does it share a common cause with them.

Of course, many actual interventions do not satisfy (a) through (c), and considerable ingenuity and hard work are often needed to ensure that the conditions are approximated in an experiment. Hence, it would be a mistake to suppose that all interventions are ideal. For instance, the Federal Reserve's decision to cut interest rates might be influenced by statistics indicating slowing economic growth, while one of the desired effects of the rate cut is to stimulate the economy. The post-intervention graph representing such a case would contain an arrow running from a variable contained in the pre-manipulation graph to the intervention, in this instance, the rate cut. Thus, the intervention in this example does not fulfill (c) listed above; the decisions of the Federal Reserve are not exogenous to the system that is the target of their interventions. Finally, it should be noted that an intervention need not come about through human activity. An intervention, as defined above, consists in a particular sort of alteration of a causal structure, whether it be brought about by deliberate action or fortuitous circumstance.



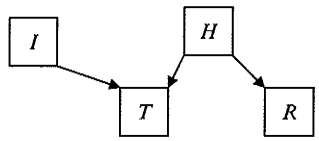


Figure 2.3 An intervention that fails (a)

The concept of an ideal intervention is of interest here primarily in virtue of its usefulness as a basis for defining other causal concepts, such as causal effect and causal relevance. It is very compelling that if X is causally relevant to Y, then ideal interventions on X alter the probability of Y, but otherwise not. For example, it is precisely this assumption that is implicit in randomized controlled experiments, which are generally regarded as the "gold standard" for assessing causal hypotheses. And it is easy to see that standard ways in which randomized controlled experiments can go wrong correspond to a failure of one or more of items (a) through (c).

For example, when some subjects in a clinical trial do not follow the experimental protocol (e.g., do not take the assigned medication as prescribed), then (a) does not obtain. This is problematic, since it allows for the possibility that there is a common cause of the variable being manipulated and the outcome. The sicker patients, for instance, might be less likely to follow the protocol, and also less likely to recover. In such circumstances, there may be a positive correlation between recovery and following the treatment protocol, even if the treatment is entirely ineffective. The point is illustrated in Figure 2.3. In the graph, T indicates treatment; H, health prior to receipt of treatment; and R, recovery.

Likewise, item (b) fails to obtain when the intervention inadvertently directly affects more than one variable in the system. In well-designed clinical trials, for example, great care is taken to ensure that both the test and the control groups are treated identically except that the former receives the treatment and the latter does not. Clearly, item (b) could fail if the intervention provided, along with the treatment, increased confidence in recovery only to those in the test group. Placebos and double-blinds are, of course, standard tactics for avoiding such difficulties. This type of failure of an intervention to be ideal could be represented graphically as in Figure 2.4. In the graph, C is some measure of the subject's confidence of recovery prior to receipt of treatment. In this example, treatment and recovery may be correlated even though the treatment itself is entirely inefficacious.

Item (c) is satisfied whenever I is the product of some purely random process. However, it may fail in the absence of randomization. For example, suppose that the researcher deliberately assigns healthier patients to the test group. Then the intervention is not exogenous, as required by (c). This situation can be represented in Figure 2.5. In such a case,

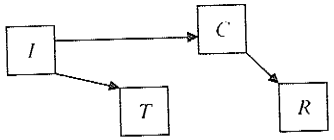


Figure 2.4 An intervention that fails (b)

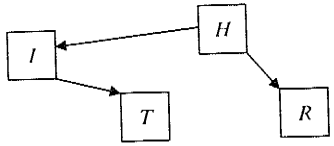


Figure 2.5 An intervention that fails (c)

treatment would be statistically associated with recovery even though the treatment is entirely ineffective.

The guiding intuition of the definitions of causal effect and causal relevance provided in what follows is that one variable, X, is causally relevant to another, Y, just in case ideal interventions on X alter the chance of Y. The philosophical aspects of this manipulationist view of causation have been discussed at length in literature on causation (cf. Woodward 2003; Hausman and Woodward 1999). Rather than recapitulate these discussions here, I make only two points. First, I do not claim that the manipulationist view of causation is the only fruitful perspective from which to approach the topic. I choose to rely upon it here because it is a useful and natural manner in which to interpret a wide range of causal claims in biology and social science. In these domains, one often wants to design interventions (e.g., therapies, policies) to achieve desired ends in complex situations in which the outcomes of such interventions cannot be predicted with certainty. Causal effects, defined in terms of the probability distribution of one variable conditional on an ideal intervention on another, are natural objects of inquiry in such circumstances.

Second, the account of causality offered in this section is not intended as a conceptual analysis of causation. A conceptual analysis of causation would consist of a necessarily true statement of the form "X causes Y if and only if...," where the ellipsis would be replaced with a Boolean combination of terms, all of which can be defined independently of the notion of causality. Since the term "ideal intervention" is itself defined by reference to causation, the account given in this section is not intended as a conceptual analysis.[3]

2.2 CAUSAL EFFECTS

The problem of extrapolation in heterogeneous populations consists in

very different from those holding in subpopulations or distinct, related populations. Thus, the notion of causal effect in a population needs to be clarified before we can get very far with our discussion. The concept of a causal effect is exemplified by controlled in experiments. In such experiments, one is interested in learning the probabilities of certain outcomes (say, recovery and non-recovery) conditional on an ideal intervention that assigns the value of some other variable, say dosage of a drug. Let X and Y be variables representing the treatment and outcome variables, respectively. I will follow the convention of using lowercase letters to denote particular values of variables represented by the same uppercase letters. For example, if X is a variable representing treatment dosage, then x is a particular dosage value, say, 200 milligrams. Pearl (2000) introduces the helpful notation $do(X = x)$, or $do(x)$ for short, to denote an ideal intervention that sets the variable X to the particular value x. Thus, the formula $P(Y \mid do(x))$ is shorthand for a function that specifies the probability distribution of Y conditional on ideal interventions that set X to any particular value x.[4] Given this notation, we can define "causal effect" in the following way (2000, 70):

> *Definition 2.2 (Causal Effect):* For any two distinct variables X and Y, the causal effect of X upon Y is $P(Y \mid do(x))$.

For example, suppose that that X and Y are each binary. Then the causal effect of X upon Y could be represented in a chart, as in Figure 2.6. Here the values of X would be set by an ideal intervention, rather than merely passively observed. Thus, where X represents treatment and Y recovery, this table represents the type of information that is desired from a randomized clinical experiment. It is important to remember that the causal effect $P(Y \mid do(x))$ need not equal the probability distribution of Y conditional on X being passively observed to have the value x, $P(Y \mid X = x)$. These conditional probability distributions may be distinct if Y is a cause of X or there are common causes of X and Y. In those cases, an ideal intervention will eliminate some causal connections between X and Y, which may thereby result in $P(Y \mid do(x))$ being distinct from $P(Y \mid X = x)$. That point is illustrated by the barometer-storm example described above.

Hitchcock (1993, 349) proposes a definition that is very similar to Pearl's, though with a few differences. In Hitchcock's version, the expression defined is not "causal effect" but "the causal relevance of the variable

X	Y=1
1	65%
0	27%

X for E," where X is a random variable and E is an event in the technical sense of set theory (i.e., a subset of the outcome space). The causal relevance of the variable X for E is then defined as $P(E \mid X = x)$, where this probability function is assumed to represent the relationship between X and E that holds when all confounding factors have been held fixed. In spite of the similarities, I shall employ Pearl's version. I prefer Pearl's definition because the expression "causal effect" has greater currency than the corresponding phrase defined by Hitchcock[5] and because Pearl's "$do(x)$" notation is very convenient.

For our purposes, one of the most important features of causal effects is that they are prone to vary according to changes in the distribution of factors in the population that affect the outcome. For instance, suppose that we are interested in the causal effect of treatment with penicillin upon recovery from streptococcal infection. This causal effect depends on, among other things, the proportion of individuals in the population who are infected with a resistant strain of the bacteria. In the extreme case in which everyone in the population is thus afflicted, treatment with penicillin may have no effect whatsoever. Since the sensitivity of causal effects to fluctuating features of populations is closely linked to the problem of extrapolation in heterogeneous populations, it is worthwhile to clarify how the term "population" is to be understood here.

Consider the statement that chemotherapy is a positive causal factor for recovery among leukemia sufferers. The most straightforward way to interpret the expression "leukemia sufferers" in this context is to take it to denote a set of real human beings who have suffered, are suffering, or will suffer from leukemia.[6] How far we intend our causal generalization to reach back into the past and extend forth into the future may vary according to several circumstances that will be considered below. The important point, however, is that the generalization is relative to some group of individuals, each of whom exists at some specific time and place in the history of the *actual* world. For example, it would be absurd to object to the proposed generalization by describing a science fiction scenario of a collection of humans whose physiology differed from those of actual people so as to reverse the effect of chemotherapy. But claims about causal effects depend on features of the actual population in other ways as well. Suppose, for instance, that the population of leukemia sufferers consists of two subgroups: one in which chemotherapy causes recovery and one in which it does the opposite. Then the overall effect of chemotherapy will depend crucially on the proportions of these two subgroups and the strength of the effect in each. Indeed, this scenario is hardly far-fetched, given the realworld variability of response to chemotherapy. Thus, I do not assume that populations are homogeneous, that they represent ideal types, or that they constitute natural kinds. It is important for the project of this book that no such assumptions be made about populations, since the problem of extrapolation in heterogeneous

Statements of causal effect, then, are typically made relative to some actual population of individuals, and the truth of the claim will generally depend on features of that population and the environment in which it is located. Unless specifically indicated otherwise, therefore, the populations of interest for the purposes of this book will be presumed to consist of individuals existing at some place and time in the actual world. The claim that chemotherapy is a positive causal factor for recovery among leukemia sufferers would ordinarily be understood to be relevant not only to people presently suffering from leukemia, but also past and future people so afflicted. However, a researcher might be more or less bold in his or her willingness to extend such a generalization into the future or past. To choose a different medical example, the effectiveness of an antibiotic in the present might not be a reliable guide to its efficacy in the future, since widespread use of it or similar antibiotics would be likely to stimulate the evolution of resistant strains. In such cases, the relevant population would be somewhat vaguely bounded in the future direction. But whether vaguely specified or not, I shall view the populations to which causal generalizations are relative as finite sets of concrete individuals located in specific environments. The individuals in the population need not be contemporaries of each other, but each must exist at some particular time and place in the history of the actual world. Subpopulations, then, are simply subsets of such sets of individuals.

2.3 CAUSAL RELEVANCE

In this section and ensuing subsections, I undertake to develop an account of the concepts of positive and negative causal relevance. Given the definition of causal effect presented above, a definition of causal relevance can be provided as follows.[7]

> *Definition 2.3 (Causal Relevance/Causal Factor):* X is causally relevant to (is a causal factor for) Y if and only if there are values y of Y and x of X such that $P(y \mid do(x)) \neq P(y)$.

This definition makes clear that the bare claim that X is causally relevant to Y is not terribly informative, since it tells us nothing about the manner of this influence. It might be that X promotes or prevents Y, or affects it in some other way. Moreover, the mere statement that X is causally relevant to Y does not tell us which values of X make a difference to Y. For example, X might have little or no effect in low dosages, but a powerful effect in higher ones.

In order to be of practical use, then, ascriptions of causal relevance generally must include some information about the manner in which the cause acts upon the effect. Expressions indicating positive or negative causal relevance are two very common ways to do this. Examples of the first sort of expression are "promotes" and "contributes to," as well as the ordinary usage of "causes." Examples of phrases that can be used to

'blocks." First, I consider the most common proposal for explicating positive and negative causal relevance, namely, the probability-raising definition. After showing why this definition is appropriate only for the case of qualitative variables, I propose a general definition that is applicable in quantitative and qualitative cases alike. Finally, I explain why an additional clause concerning contextual unanimity should not be added to the definition.

2.3.1 The Probability-Raising Definition

Let us begin with the most common probabilistic rendering of such expressions as "A promotes B" or "A contributes to B," namely, the *probability-raising definition* (cf. Suppes 1970; Eells 1991). In such approaches, A and B would typically be interpreted as events or propositions, rather than quantitative variables. Suppose that C represents all common causes of A and B. Then such theories typically assert that A is a positive causal factor for B just in case A is temporally prior to B, and $P(B \mid A \,\&\, C)$ is greater than $P(B \mid \neg A \,\&\, C)$. The probability-raising definition is reasonable in simple examples in which the causes are represented by binary variables, as in classic treatment/non-treatment clinical experiments, but is less adequate when applied to examples involving quantitative variables. The difficulty in question arises in the form of the "problem of disjunctive factors" (cf. Humphreys 1989, 40–41; Eells 1991, 144–68; Hitchcock 1993).

The problem is that when A represents a quantity, the negation of A, $\neg A$, indicates a disjunction of possible values $\{A_1, \ldots, A_n\}$. Moreover, whether $P(B \mid A \,\&\, C)$ is greater than, equal to, or less than $P(B \mid \neg A \,\&\, C)$ can depend on the probabilities $P(A_1), \ldots, P(A_n)$. Consider the following example due to Paul Humphreys (1989, 40–41). We are concerned to test the effectiveness of treatment with a particular drug, A, in bringing about recovery, B. We conduct a randomized controlled experiment in which the subjects are divided into three groups. The first group receives a placebo (A_0); the second, a moderate dose of the drug (A_1); and the third, a large dose (A_2). Suppose that the probabilities in the experiment are the following: $P(B \mid A_0) = .2$, $P(B \mid A_1) = .4$, and $P(B \mid A_2) = .9$. Then, given that $P(A_0) = P(A_1) = P(A_2) = 1/3$, we have $P(B \mid \neg A_1) = .55 > P(B \mid A_1)$. Hence, according the probability-raising definition, moderate doses of the drug prevent recovery. However, this result is quite odd, since the probability of recovery with moderate doses is greater than with a placebo. Moreover, in the example, whether A_1 raises or lowers the probability of B depends on the relative frequency of the treatment assignments. For instance, if $P(A_0) = 7/12$ and $P(A_2) = 1/12$ while all of the other numbers in the example remain the same, then A_1 raises the probability of B.

In general, the problem of disjunctive factors is motivated by the idea that whether X is positively or negatively relevant to Y should not depend upon how frequently particular values of X happen to occur.

are intended to provide information concerning the effects of interventions. If X promotes Y, then increasing X ought to be an effective strategy for increasing Y. But in order for claims concerning positive causal relevance to play this role, it is important that they be invariant under interventions. Furthermore, it is obvious that an intervention normally will alter the probability distribution of the cause, since an intervention seeks to change the distribution of the effect by changing the distribution of the cause. For example, a government health initiative might attempt to reduce the prevalence of lung cancer by reducing the frequency of smoking. Hence, if claims about positive and negative causal relevance are to provide useful guidance concerning the outcomes of interventions, they should be invariant under changes to probability distribution of the cause.

In addition to posing a difficulty for the probability-raising definition of positive causal relevance, the problem of disjunctive factors also provides an argument against using correlation as a measure of positive and negative causal relevance. The correlation between X and Y, $\rho(X, Y)$, is defined as follows:

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\ \text{var}(Y)}}$$

In this equation, $\text{cov}(X, Y)$ is the covariance of X and Y, which is equal to $E(XY) - E(X)E(Y)$,[8] while $\text{var}(X)$ is the variance of X, which equals $E((X - E(X))^2)$. As long as the values of X and Y consist solely of real numbers, the denominator of the right-hand side of the above equation is equal to or greater than zero. If it is also the case that neither variable is constant (i.e., there is some variation in both X and Y), then the denominator is strictly positive. Hence, for all situations that need concern us here, covariance determines whether the correlation is positive or negative. Yet whether the covariance is positive or negative can depend upon the probability distribution of the cause. For example, consider a case in which X and Y each have three possible values: 0, 1, and 2. Suppose, moreover, that the values of X and Y tend to coincide when $X = 2$, but tend to differ when $X = 1$. In such a case, whether the overall correlation is positive or negative may depend upon the probabilities $P(X = 2)$ and $P(X = 1)$.[9]

One way to deal with problem of disjunctive factors is to propose that attributions of causal relevance are always, though sometimes implicitly, comparisons between probabilities conditional on particular values of the cause (cf. Humphreys 1989; Holland 1986).[10] Hence, in the above example, we could say that moderate doses promote recovery because $P(B|A_0) < P(B|A_1)$. This proposal is quite sensible when the cause is a nonbinary, qualitative variable. For example, suppose one is interested in the influence of race upon employment, where employment is treated as a binary variable and race is treated as a qualitative variable that can take more than two values, say, white, black, Hispanic, or Asian. Suppose

hat the rate of employment among Asians is highest of all, and that of whites is higher than for both blacks and Hispanics. In this case, whether being white is positively relevant to employment depends on the proportions of the distinct races in the population, in direct analogy to Humphreys's example. It is plausible in this case to insist that claims about positive causal relevance are inherently contrastive, so that the claim that being white is a positive causal factor for employment is understood, implicitly or explicitly, in contrast to being black or Hispanic.

For quantitative variables, however, it is unreasonable to insist that attributions of causal relevance must always be relative to a pair of comparative values of the cause. For example, the claim that smoking causes lung cancer is not equivalent to the statement that, say, that the probability of cancer is greater if you smoke two packs a day rather than just one. The statement that smoking causes lung cancer entails something *in general* about the relationship between cigarette smoking and cancer. A general claim of this sort cannot be identified with a claim about the effects of smoking a specific number of cigarettes any more than the claim that all sparrows have wings can be equated with a claim about a particular bird. Notice that a difference between the smoking example and cases involving qualitative variables (such as the race-employment example) is that it is sensible to speak of the consequences of *increasing* or *decreasing* the cause. The number of cigarettes smoked per day may be raised or lowered; yet it would be nonsensical to speak of increasing or decreasing a person's race.

The problem of disjunctive causal factors suggests that comparisons of the probability of the effect, given specific values of a cause, are at best an adequate account of causal relevance for qualitative variables. That leaves us with the problem of explaining what expressions such as "X inhibits Y" or "X promotes Y" mean when X or Y is a quantitative variable.

2.3.2 Causal Relevance for Quantitative Variables

Stated in terms of the concepts presented above, Hitchcock's (1993) proposal is that claims about positive and negative causal relevance provide qualitative information about causal effects. But what qualitative information, exactly? Consider expressions that indicate positive causal relevance, such as "X promotes Y" or "X causes Y." How should such statements be understood when X and Y are quantitative variables? An appealing idea is that such claims are understood to mean that *increases* in X produce *increases* in Y. Likewise, "X prevents Y" means that *increases* in X produce *decreases* in Y. Hitchcock's discussion of the smoking-cancer example (1995, 261–62) suggests that he, too, shares this intuition.[11] However, this intuitive idea can be interpreted in more than one way. Both interpretations involve reference to some interval of values of the cause. On what I call the *comparative* interpretation, the claim of positive causal relevance indicates that the value of Y is greater when X is raised from

to what I term the *monotonic* interpretation, positive causal relevance means that the value of Y increases monotonically with X throughout the interval. The cause might be positively relevant in one of these senses but not the other.

Some preliminary clarification is required to explore these ideas in the present context. In particular, the intuitive idea that positive causal relevance means that increases in the cause yield increases in the effect requires modification for cases in which the relationship between cause and effect is probabilistic. For in that case, increases in the cause do not *always* produce increases in the effect. However, the intuition is naturally extended to probabilistic examples as follows: increases in the cause yield increases in the *expected value* of the effect. Here "expected value" can be understood by means of the notion of an *average causal effect* or, in symbols, $E(Y \mid do(x))$. In the case in which Y is discrete, $E(Y \mid do(x)) = \sum_y yP(y \mid do(x))$. When Y is continuous, $E(Y \mid do(x)) = \int_{-\infty}^{+\infty} yg(y \mid do(x))dx$, where $g(y \mid do(x))$ is a probability density function defined as $P(a \leq Y \leq b \mid do(x)) = \int_a^b g(y \mid do(x))dx$, for any pair of real numbers a and b.

Notice that the average causal effect omits all information concerning the variance of Y, which is an important point, since interventions on X might alter the variance of Y without changing its expected value. For example, imagine a social program that redistributes wealth from rich to poor. Such a program would clearly affect the distribution of wealth in the society but could leave the average, or expected, wealth unchanged. Although the program is causally relevant to wealth, it would be odd to say that the program promotes or inhibits it. However, it would be natural to say that the program promotes economic equality, a thought which is easily accommodated by the present proposal.[12] Thus, I suggest that claims about positive and negative causal relevance are insensitive to changes in the distribution of the effect that leave its mean unaltered. That of course is not to deny that it is important in some circumstances to know how X affects the distribution of Y aside from changing its mean. Rather, the point is merely that such information is not conveyed by claims about positive and negative causal relevance. Note that this situation can arise only if the variable representing the effect is not binary. Given that discussions of causal relevance have tended to focus on binary events or properties, it is not surprising that this complication has not been discussed.

Given these preliminaries, the comparative and monotonic interpretations of positive relevance can be stated more precisely. Let ϑ be an interval of values of X, and let x_0 be some appropriate comparative value of X such that, for every x in ϑ, $x_0 < x$. Then X is a *comparative positive causal factor* for Y within the interval ϑ of X if and only if, for all x in ϑ, $E(Y \mid do(x)) > E(Y \mid do(x_0))$. In contrast, X is a *monotonic positive causal factor* for Y if and only if, for all x in ϑ, $\frac{\partial}{\partial x}E(Y \mid do(x)) > 0$. In other words, X is a monotonic positive causal factor for Y within ϑ when the function $E(Y$

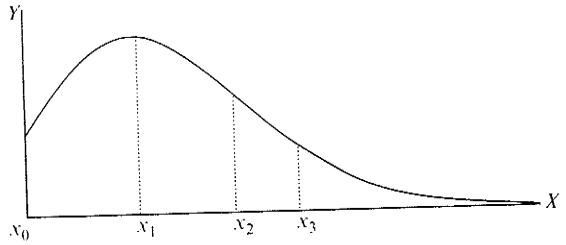


Figure 2.7 The fertilizer example

$| do(x))$ increases throughout ϑ. Definitions of comparative and monotonic *negative* causal relevance can be obtained by substituting "<" for ">" in these two definitions. Likewise, definitions of comparatively and monotonically *neutral* can be obtained by substituting "=" for ">" in the same places.[13]

It will be helpful to illustrate these definitions with a concrete example. Suppose that we are interested in the effect of a certain fertilizer on the growth of a particular species of plant. Let the variable X be a measure of the dosage of the fertilizer and Y a measure of the height of the plant. Imagine that the average causal effect is represented by the curve in Figure 2.7. In the figure, $E(Y \mid do(x))$ increases from x_0 to x_1, where it reaches its maximum; thereafter, $E(Y \mid do(x))$ decreases and ultimately converges to zero.[14] Hence, if x_0 is our comparative value, then X is both a comparative and monotonic positive causal factor for Y within the interval (x_0, x_1). Similarly, X is both a comparative and monotonic negative causal factor for Y in any interval to the right-hand side of x_3. But within the interval (x_0, x_2), X is a comparative positive causal factor for Y, but neither a positive, negative, nor neutral monotonic factor. And within the interval (x_1, x_2), X is a comparative positive causal factor for Y but a monotonic negative factor.

It would be cumbersome and inconvenient, however, to operate with two definitions of positive relevance throughout the remainder of this book. In what follows, I will say that X is a positive causal factor for Y (full stop) exactly if X is both a comparative and a monotonic causal factor for Y.

> *Definition 2.4 (Positive Causal Relevance):* Let ϑ be an interval of values of X, and let x_0 be some appropriate comparative value of X such that, for every x in ϑ, $x_0 < x$. Then X is a *positive causal factor* for Y within the interval ϑ of X if and only if X is both a comparative and a monotonic positive causal factor for Y within the interval ϑ.

The definition of negative causal relevance can be obtained from Definition 2.4 through a simple reversal of inequalities, as explained above. So letting $X = 0$ be our comparative value in the fertilizer example,

open interval (x_0, x_1). Likewise, X is a negative causal factor for Y within the interval x_3 to infinity. Thus, Definition 2.4 coincides with the natural judgment that the fertilizer promotes growth in moderate doses but has the opposite effect in very large doses. Definition 2.4 also nicely treats Humphreys's example that served to illustrate the problem of disjunctive factors. Let Y be a binary variable representing recovery, and let X be a continuous variable representing the treatment dosage. Letting $Y = 1$ stand for recovery and $Y = 0$ for non-recovery, the average causal effect, $E(Y \mid do(x))$, is equal to $P(Y = 1 \mid do(x))$. In Humphreys's example, we know $P(Y = 1 \mid do(x))$ for three values of X, ranging from a zero dosage to a large one. These probabilities suggest that $P(Y = 1 \mid do(x))$ increases monotonically, at least for doses no greater than the largest administered in the experiment. Given that this inference is correct, Definition 2.4 entails that X is a positive causal factor for Y in the interval $(x_0, x_2]$, where x_0 is the zero dosage and x_2 is the large one.

Definition 2.4, then, can be regarded as delineating unambiguous cases of positive causal relevance among quantitative variables. For instance, although the fertilizer is clearly a positive causal factor for growth in moderate doses and a negative factor in very large doses, it is unclear how the effect of intermediate doses should be characterized. For example, when ϑ is the interval (x_1, x_2) in Figure 2.7, $E(Y \mid do(x))$ is greater than $E(Y \mid do(x_0))$ for all x in ϑ, while for all x in ϑ, $\frac{\partial}{\partial x} E(Y \mid do(x))$ is negative. Hence, Definition 2.4 indicates that the fertilizer is neither a positive, a negative, nor a neutral factor for growth in the interval (x_1, x_2), which is a way of indicating the ambiguous nature of the situation. Moreover, given Definition 2.4, any proposition demonstrated concerning conditions in which claims about positive causal relevance can be extrapolated automatically holds for both the comparative and the monotonic senses of that notion. Of course, the practical convenience of Definition 2.4 for the purposes of this book does not show that it represents the one true way to understand positive relevance for quantitative variables. There might be contexts wherein causal relevance is most naturally understood in terms of either comparative or monotonic relevance alone. Nevertheless, I think Definition 2.4 is a reasonable compromise for the present purposes.

An additional nice feature of Definition 2.4 is that it enables us to treat negative and positive causal relevance for qualitative variables as a special case simply by disregarding the monotonicity condition, which is clearly inapplicable in the qualitative case, and by having ϑ be a single value of X rather than an interval. For instance, when X is binary, $x_0 = 0$ and $\vartheta = [1]$ (since $x_0 < x$, for all $x \in \vartheta$). When X is a qualitative variable with more than two possible values (as in the race-employment example), the definition simplifies to the proposal considered in section 2.3.1 that claims about causal relevance are always, explicitly or implicitly, comparisons involving two values of the cause. Thus, the probability-raising definition of causal relevance is a special case of the definition just presented, namely, the case in which X is a qualitative variable. Hence, any propos-

ition that is true of positive, neutral, and negative factors in the case of quantitative variables is also true for qualitative variables (though not vice versa).

Definition 2.4 also captures Hitchcock's (1993, 2003) insight that positive and negative causal relevance are merely two varieties among many. In Definition 2.4, positive, negative, and neutral causal relevance are not collectively exhaustive. For instance, as explained above, in Figure 2.7, X is not neutral with respect to Y in the interval (x_1, x_2), but neither is it a positively nor a negatively relevant causal factor. Note that situations of this kind can arise only if the function $E(Y \mid do(x))$ is nonmonotonic. When $E(Y \mid do(x))$ is constant, monotonically increasing, or decreasing, comparative and monotonic relevance are equivalent.

When $E(Y \mid do(x))$ is a nonmonotonic function, the language of positive and negative causal factors can still be useful (as the fertilizer example illustrates), but may be incapable of describing important aspects of the average causal effect. For example, it is useful to know the value of X for which the function represented in Figure 2.7 reaches its maximum, since this represents the optimum dosage. But this information cannot be expressed in the language of positive and negative causal factors. A similar point can be made with respect to some monotonic cases. For example, suppose that $E(Y \mid do(x))$ increases monotonically and asymptotically converges to the value n. Then it may be important to know the value of n and how quickly $E(Y \mid do(x))$ converges to it, yet such information cannot be expressed in terms of positive and negative causal relevance. In short, the language of positive and negative causal relevance can convey useful information even with regard to nonmonotonic curves, but the more complex the shape of the curve, the more likely that it will be expedient to supplement, or perhaps even replace, talk of positive and negative causal relevance with more detailed descriptions of the shape of $E(Y \mid do(x))$.

The phrase "appropriate comparative value" in Definition 2.4 requires some further comment. In many cases it is very natural to let x_0 be 0; for example, a zero dosage of fertilizer. However, I do not insist that there is one objectively correct choice of the comparative value x_0 in each case. Claims to the effect that X is a positive causal factor for Y serve to provide qualitative information about $E(Y \mid do(x))$. We may wish to convey different sorts of information about the same function in different contexts, and different choices of x_0 may sometimes be useful for this purpose. Notice, however, that in the monotonic case, it makes no difference which value of X we choose for the comparison. For some nonmonotonic functions, the choice of x_0 may be highly arbitrary as well as very relevant to whether X is a positive causal factor for Y, according to Definition 2.4. In such cases, I suggest that the language of positive and negative causal factors is of limited utility.

The only constraint placed on x_0 in Definition 2.4 is that it be strictly less than each point in the interval ϑ.[15] However, there are cases in which it is not implausible that the comparative value x_0 would be greater than

> We can imagine a country in which almost everyone smokes two packs per day, and in which the surgeon general admonishes citizens to cut back to one pack per day. In such a context, it might be natural to say that smoking (only) one pack per day inhibits lung cancer.... (1995, 262)

In this example, the comparative point is two packs of cigarettes per day, while the interval is one pack per day or less. I agree that, in the imagined context, such a choice of interval and comparison point might be convenient for conveying the information that reducing the number of cigarettes smoked from two packs a day to just one reduces the chance of lung cancer. But does this mean that one should say in Hitchcock's example that smoking *prevents* lung cancer? Let us consider what Definition 2.4 has to say about this case.

Observe that Definition 2.4 is not applicable in the case in which the comparative value x_0 is greater than every member of ϑ. Let us consider, then, a modified version of Definition 2.4 in which the comparison point x_0 is greater than every member of the interval. This has the effect of putting the causal claim in terms of the effect that *decreasing* X has upon the expected value of Y. In Hitchcock's example, the envisioned *decrease* in smoking would be expected to produce a corresponding *decrease* in the prevalence of lung cancer. The original version of Definition 2.4, on the other hand, is designed for cases in which causal claims are expressed in terms of the consequences of *increases* in the independent variable. These two modes of expression convey the same information, since decreases in X produce decreases in Y just in case increases in X produce increases in Y. Given that the same information about the function $E(Y \mid do(x))$ is being communicated in both cases, it would be quite misleading indeed to label X's influence upon Y "negative" in one case and "positive" in the other. Thus, if we are to modify Definition 2.4 so that the comparison value may be greater than the values in the interval, we should also reverse the inequality in the definition of comparative causal relevance. Thus modified, Definition 2.4 would state that smoking is a positive causal factor for lung cancer in Hitchcock's example.

With the definition of positive and negative causal relevance, three types of causal claims have been described. In descending order of the precision of the information provided by each type, we have: causal effects, average causal effects, and claims concerning causal relevance. Causal effects and average causal effects can be estimated in some contexts, but are extremely sensitive to changes in background conditions. Consequently, qualitative claims about positive and negative causal relevance are useful in that their roughness and imprecision make them less dependent on the particular circumstances of a specific population. Although it is extremely unlikely that the causal effect found in one heterogeneous population is exactly replicated in another, it may be reasonable to expect that a positive causal factor in one population is also such in other related populations.

2.3.3 Contextual Unanimity

It is sometimes insisted that claims about causal relevance can be properly made only with respect to populations that satisfy a condition called *contextual unanimity* (cf. Cartwright 1983; Eells and Sober 1983; Eells 1986, 1987, 1991). Contextual unanimity obtains when the positive causal factor is such not merely for the population as a whole, but also for every subset of it.[16] However, I shall *not* include contextual unanimity as a part of the definition of positive and negative causal relevance.

Writing contextual unanimity into the definition would make it very hard to see how positive causal relevance could be discovered by the usual scientific means designed for such purposes, particularly randomized controlled experiments (cf. Dupré 1993, 200–1). A randomized controlled experiment may tell us that the cause is positively relevant in the population overall, but such a result is consistent with that effect being neutralized or even reversed in subpopulations. Indeed, among heterogeneous populations it is quite common that there are unknown factors capable of disrupting the mechanism linking cause and effect. Consequently, if contextual unanimity is part of the meaning of claims concerning positive causal relevance, then it is unclear how one could establish that smoking causes cancer, HIV causes AIDS, and so on. In short, if a definition of positive and negative causal relevance is to be applicable to typical examples in biology, medicine, and social science, then it is inevitable that it must allow such claims to be made with respect to heterogeneous populations in which the overall causal effect may be nullified or even reversed in subpopulations. Since claims of positive and negative causal relevance *are* frequently made with respect to heterogeneous populations, it is quite implausible that contextual unanimity is inherent in the meaning of such claims.

Contextual unanimity is best viewed not as a part of the *meaning* of claims concerning positive causal relevance, but as a circumstance that may facilitate extrapolation if present. Adding contextual unanimity to the *definition* of causal relevance is not a fruitful strategy with respect to extrapolation for two reasons. First, as noted above, such an addition would make it practically impossible to learn causal relevance relationships in many areas of biology and social science. Second, although the satisfaction of contextual unanimity can aid extrapolation, it is neither *necessary* nor *sufficient* in general for this purpose.

Extrapolation can be possible even when contextual unanimity does not obtain. In fact, Chapter 6 examines several circumstances that suffice for extrapolating claims about positive causal relevance, *none* of which require contextual unanimity. Consider one very simple example. Imagine a vaccine that is known to be effective in the general population P, although there are some rare cases in which the vaccine has the opposite effect of what is intended. Clearly, contextual unanimity does

want to know whether the vaccine also inhibits infection in P′. In spite of the failure of contextual unanimity, we would be able to conclude that the vaccine is effective in P′ if we knew that the proportion of negative and positive reactions to the vaccine in P′ is similar to that of the general population P.[17]

Contextual unanimity is also not always *sufficient* for extrapolation. This is most obviously the case when one wishes to extrapolate quantitative information concerning the causal effect, information that may be of practical significance. Even if positive contextual unanimity obtains, for example, the cause may have a strong effect in some populations and a minuscule effect in others. Moreover, there may also be qualitative features of the causal effect that are not expressible in terms of negative and positive causal relevance. For instance, suppose that $E(Y \mid do(x))$ increases monotonically and asymptotically converges to the value n. The value of n, and how quickly the function converges to it, may be important information. Yet even if the population is contextually unanimous, the value of n and the rate of convergence in the population as a whole may differ markedly from that in some subpopulations. This is an extrapolation problem that contextual unanimity, even if it were an available assumption, would not suffice to resolve.

Contextual unanimity is not the only circumstance that might facilitate extrapolation in some circumstances. For example, Chapter 6 examines a condition I call *consonance* that, put roughly, requires that there not be counteracting causal paths from cause to effect. Contextual unanimity and related conditions, such as consonance, should not be viewed as part of the meaning of claims about causal relevance. Instead, they should be regarded as premises that can aid extrapolation in certain types of cases, though not necessarily others. Although I suspect that contextual unanimity is very rarely a justifiable assumption in interesting biological or social science examples, I think that consonance is reasonable in some circumstances. In Chapter 6, I explain consonance in greater detail, consider the circumstances under which it is a reasonable assumption, and examine how it facilitates extrapolating claims concerning positive or negative causal relevance.

2.4 CONCLUSION

Heterogeneity poses a challenge for extrapolation because it raises the possibility that a causal effect in one population might differ in some significant respect from that found in other, related populations. Consequently, clear definitions of "causal effect" and of common expressions for indicating qualitative features of causal effects—particularly, positive and negative causal relevance—need to be given before much progress regarding this problem can be made. This chapter has endeavored to provide these definitions. Let us turn, then, to a consideration of the relation between these probabilistic causal concepts and mechanisms.

3

Causal Structure and Mechanisms

An important prerequisite for exploring the mechanisms approach to extrapolation is to explain what the qualitative concept of a mechanism has to do with probabilistic causal concepts such as causal effect and causal relevance. That is the task undertaken in this chapter and the next. In this chapter, I attempt to show that, for a broad range of cases of interest to the present study, it is reasonable to identify mechanisms with what is called *causal structure* in work on the problem of inferring causal conclusions from statistical data (cf. Glymour and Cooper 1999; Spirtes, Glymour, and Scheines 2000; Pearl 2000; Neopolitan 2004). Accomplishing this necessitates saying something about what causal structure is, and when and why mechanisms can be identified with it.

Explaining how this works involves reconsidering the manner in which analytic philosophers have traditionally approached the topic of causality. One of the primary activities (and perhaps *the* primary activity) of traditional analytic philosophy is conceptual analysis. I understand conceptual analysis to consist of providing necessary and sufficient conditions for the application of an interesting yet somewhat unclear term (e.g., "explanation," "cause"), where these conditions satisfy the following two properties. First, the conditions are stated via concepts that can be defined independently of the target of the definition. Second, the usage of the term recommended by the analysis must agree tolerably well with the intuitions of native speakers in all conceivable circumstances. However, conceptual analysis has decidedly fallen from favor in recent years in the philosophy of science. For example, leading accounts of causality in the recent philosophy of science literature (cf. Hausman 1998; Dowe 2000; Woodward 2003) explicitly disavow any intention to provide a conceptual analysis in the sense just described. Rather than conceptual analysis, these authors endeavor to develop an account of causality that is informed by current scientific theories and methodology. Dowe, whose approach to causation owes much to Wesley Salmon (1984), strives for what he terms an *empirical analysis* of causality, that is, "to discover what causation is in the objective world" (Dowe 2000, 1). Dowe regards current physical theory as the most reliable source of information that would serve as a basis of an answer to this question.

But there is a simple objection to any program that would proceed with empirical analysis before conceptual analysis is complete: without prior conceptual analysis it is unclear what basis there is for asserting that the

from ordinary language. David Lewis has posed this objection in the context of a discussion of the philosophy of mind, but it transfers easily to discussions of causation. In Lewis's words:

> Arbiters of fashion proclaim that analysis is out of date. Yet without it, I see no possible way to establish that any feature of the world does or does not deserve a name drawn from our traditional mental vocabulary. (1994, 415)

After considering and rejecting Dowe's response to this objection, I propose that a better answer derives from the view that causal locutions should be treated as theoretical terms in the sense of the Ramsey-Lewis account, according to which theoretical terms are a kind of definite description (cf. Lewis 1970). Given this perspective, an empirical analysis should be based upon a meaning postulate that specifies a particular role associated with the term in question. I will concentrate on two roles ascribed to causal structure; in particular, causal structure is that which generates probability distributions and indicates how these distributions change given interventions.

From this starting point, an empirical analysis of causal structure consists of indicating what fulfills these roles in a particular domain. Making the case for identifying mechanisms with causal structure requires some general argument for supposing that mechanisms are modular, in the sense that it is possible to alter one component without disrupting the functioning of the others. I explain how evolutionary theory can support the claim that modularity is likely to be a pervasive feature of mechanisms. However, this argument is, at present, on firmer ground in molecular biology than in social science, making the motivation for identifying causal structure with mechanisms somewhat more tentative in the latter case. An implication of this discussion is that empirical analyses of causation depend on domain-specific scientific details and hence may differ for distinct phenomena. The question of whether social mechanisms should be identified with causal structure, and under what circumstances, will be explored in further detail in Chapter 8.

3.1 IT'S NICE, BUT IS IT CAUSALITY?

An empirical analysis of causation proceeds by examining the question of what causation is in the world. For example, Dowe's conserved quantity theory advances the following two propositions as the foundation of an answer to that question:

> *CQ1.* A *causal process* is a world line of an object that possesses a conserved quantity.
>
> *CQ2.* A *causal interaction* is an intersection of world lines that involves exchange of a conserved quantity. (2000, 90)

It is striking how removed this analysis is from many ordinary discussions of causation. For instance, it is unclear what relevance exchanges of conserved quantities have to the claim that the vitamin C tablets that Bob ate did not cause him to recover from his cold.[1]

Lewis's objection, then, seems quite apt: the conserved quantity theory is interesting, but why should one regard it as an account of causality? And how can this question be answered without presupposing a conceptual analysis? Dowe responds to this objection in the following way:

> In drawing explicitly on scientific judgments rather than on intuitions about how we use the word, we nevertheless automatically connect to our everyday concept to some extent, because the word cause as scientists use it in those scientific situations must make some historical or genealogical connection to everyday language. (2000, 9)

Thus, basing an analysis of causation on current science connects to commonsense ideas concerning the meaning of "cause" since the usage of the term by scientists is linked to that of ordinary folk. But does this mean that empirical analysis simply amounts to a conceptual analysis of scientists' concept of causation? Dowe makes it clear that this is not his intent: "The task of empirical analysis . . . is not a conceptual analysis of scientists' usage of a term" (2000, 10). Rather, he maintains that the empirical analysis he pursues aims to explicate the concept of causation "implicit in scientific theories" (2000, 11).

The main difficulty I see with this response is that it is highly questionable whether there *is* a concept of causation implicit in current scientific theory. As Dowe observes, no physical theory contains "cause" as an explicitly defined term (2000, 9), and consequently any proposed empirical analysis of causation must inevitably be a substantive thesis over and above what is given by science (Bontly 2006, 182–83). Moreover, there are several ways that one could interpret causation in the light of current science, and it seems unavoidable that arguments for choosing one approach over another will appeal to intuitions about the proper usage of the word "cause." To take just one issue, consider whether causation requires determinism. Dowe argues that the answer is *no*, on the grounds of an example concerning exposure to radioactive material.

> If I bring a bucket of Pb^{210} into the room, and you get radiation sickness, then doubtless I am responsible for your ailment. But in this type of case, I cannot be morally responsible for an action for which I am not causally responsible. (2000, 23)

Thus, given the scientifically plausible assumption that the decay of Pb^{210} is a fundamentally indeterministic process, it follows that indeterministic causation exists.

Although the above argument is interesting and perhaps even persua-

implicit in physical theory. Dowe's argument depends crucially on the thesis that moral responsibility (at least in some unspecified class of cases of which the present one is an example) entails causal influence. But what is the basis of any such principle linking moral responsibility and causation? Surely it is not physical theory. Rather, any grounding for it would reside in the interconnection of ordinary concepts of responsibility and causality. As a result, one who maintained that determinism is a fundamental aspect of the concept of causality (e.g., Pearl 2000, 26–27) could avoid the conclusion of Dowe's argument by rejecting the claim that moral responsibility implies causal influence. For example, I might have a moral responsibility to provide assistance to starving people in a distant land despite the fact that I am in no way causally responsible for their unfortunate situation. Thus, Dowe's use of current physics to argue for indeterministic causation requires an antecedent clarification of the relationship between causation and moral responsibility.

Physical theory certainly does have implications for the nature of causation. In the foregoing example, modern physics makes it difficult to maintain both that causation is inherently tied to determinism and that moral responsibility entails causal influence. But this does not show that there is a single account of causality implicit in physical theory, since several different accounts of causation can be made consistent with modern science, depending on what position one takes regarding the interconnections between causation and such things as responsibility, human agency, determinism, temporal priority, spatiotemporal contiguity, and so on. Yet one significant aim of conceptual analysis is to settle questions concerning such interconnections. Hence, we are led straight back to Lewis's objection: empirical analysis cannot fruitfully proceed until matters of conceptual analysis have been settled.

Let us consider a different account of how an empirical analysis of causation can proceed even in the absence of a successfully completed conceptual analysis.

3.2 CAUSALITY AND THEORETICAL TERMS

In this section, I suggest that the cogency of empirical analysis without a successfully completed conceptual analysis can be defended by considering causal locutions as theoretical terms in the sense of the Ramsey-Lewis account (Ramsey 1954; Lewis 1970). The Ramsey-Lewis account proposes to treat theoretical terms as a type of definite description stated via antecedently understood concepts:[2] the theoretical entity is simply that (if anything) which satisfies the description. For example, in eighteenth-century chemistry, phlogiston is that which is present in all flammable objects and is emitted during the process of combustion. In Lavoisier's chemistry, oxygen is that which is absorbed during combustion and is necessary for the formation of acids.

Several authors have suggested that the Ramsey-Lewis account, in addition to applying to deliberately introduced terms of scientific theories, could also be appropriate with regard to concepts falling more squarely in the province of philosophy. For example, Michael Tooley (1987) and Peter Menzies (1996) take such an approach to causation, and Dowe (2000, 49–51) sympathetically considers the idea with respect to transference theories of causation.[3] In Dowe's formulation, such an analysis of causation would consist of three components: a meaning postulate, a contingent hypothesis, and an a posteriori identity (2000, 49). The meaning postulate is the definite description that specifies some important feature of causation: causality is that which does __. For example, one plausible claim is that causation is that which underlies the possibility of predicting the consequences of interventions (cf. Menzies and Price 1993; Woodward 2003). The contingent hypothesis would then be an empirical claim about what things in the world fulfill this role in a given domain, while the a posteriori identification would assert that (in the domain in question) causation is identical to the entity or process indicated in the contingent hypothesis.

The question, then, is how to decide what the meaning postulate should be. An agreed-upon conceptual analysis, if one were available, clearly would be one possible basis for answering this question. For example, Tooley treats his proposal regarding the meaning postulate as a conceptual analysis (cf. Tooley 1987, 25–28). If this were the only possible way to justify one's choice of meaning postulate, then Lewis's argument that empirical analysis cannot proceed until matters of conceptual analysis have been settled would be vindicated. But there is another possibility: the meaning postulate could be derived from empirical observations of the use of causal language. For example, Thomas Bontly proposes that we regard "the concept of causation as a concept defined by its place in an inferential system or network, by the inferences it licenses and those that license it" (2006, 191). Given this perspective, the meaning postulate should be based on inferences that people actually make to and from causation. A meaning postulate, then, should indicate something that is generally regarded as evidence for causal claims as well as something that is judged to be a consequence of causal claims. A meaning postulate that focuses on the connection between causation and predicting the outcomes of interventions does both of these things. The connection between causal claims and effective strategies for achieving ends has been emphasized by many authors (cf. Cartwright 1983, chap. 1; Mellor 1988, 230; Hoover 2001; Woodward 2003). Moreover, carefully controlled interventions are generally regarded as the most reliable scientific means for testing causal claims. There is also experimental evidence that preschool-age children regard interventions as an especially effective way of learning what causes what (Kushnir and Gopnik 2005). Similarly, covariance is generally regarded as a consequence of causal relationships and

Thus, either manipulation or covariance of the right sort is a potential basis for a meaning postulate in an empirical analysis of causation. In fact, the meaning postulate that will be discussed below—according to which causal structure is that which generates probability distributions and provides information about how they change under interventions—combines both notions. Physical contiguity is a third factor that is often relevant to causal inferences, and it is presumably the guiding thought behind Dowe's conserved quantity theory. However, physical contiguity alone is rarely sufficient to infer causation, since one event might be physically adjacent to another without having caused it. Not surprisingly, in his definition of "C causes E," Dowe combines the definitions of causal process and interaction presented above with a requirement that the cause raise the chance of the effect (2000, 167).

The link between causation and manipulation is doubtful as a conceptual analysis of causation, since specifying what a manipulation or intervention is will inevitably involve references to causation. Nevertheless, a principle linking causation to manipulation can serve as an appropriate meaning postulate for an empirical analysis that treats causation as a theoretical term in the sense of the Ramsey-Lewis theory. If it can be shown that the feature of the world specified in the empirical analysis makes effective manipulation possible, then there is a straightforward answer to the question: Why call it causation? Whatever causation is, knowledge of it is often important for indicating effective and ineffective strategies for achieving ends. Hence, if one identified a general feature of the world that fulfilled this function, then one would have a legitimate claim to be describing causation.

It may be objected that the connection between manipulation and causation could not serve as a meaning postulate, since manipulation is a causal concept, whereas the terms in the meaning postulate are supposed to be antecedently understood. In response, I claim that manipulation and intervention are antecedently understood: they are drawn from the vocabulary of ordinary English and everyday life. (Of course, that does not preclude the usefulness of introducing a framework for discussing them more clearly, as done in section 2.1.) The key point is that *antecedently understood* is a criterion distinct from *independently definable*: we have a reasonably clear idea of what an intervention is, regardless of whether we can define the term in a manner that eschews all reference to causation. Consequently, it is legitimate to use intervention as the basis of the meaning postulate for a Ramsey-Lewis-style definition of "causal structure."

Another possible objection is that without a conceptual analysis of causation, there will be several potential starting points for an empirical analysis of causation. I think it is quite right that there may be several reasonable choices for starting points for an empirical analysis of causation, and that different starting points might lead to separate destinations. However, this is a problem only if one supposes that there must be

a monolithic concept of causation for which a unique empirical analysis must be given. In contrast, I see no reason to rule out at the start of inquiry the possibility that the notion of causation is multifaceted.[4] Given the account proposed here, empirical analyses of causation might be pluralistic in two ways. First, a single meaning postulate might be realized differently in distinct domains. For instance, that which generates probability distributions and provides information about how they change under interventions might be one kind of thing in fundamental physics and another in molecular biology and something else again in economics. Second, there may be several reasonable meaning postulates that lead to distinct empirical analyses even within the same domain of inquiry. For example, an empirical analysis based on manipulation might lead to results different from one that emphasizes physical contiguity. The potential for this second type of pluralism raises the question of whether there are common threads linking the several meaning postulates, or whether "causation" is simply an ambiguous term with several distinct meanings. My own view is that the various causal concepts are all closely linked elements of a network of concepts relating to practical reason. However, the account of extrapolation developed in this book does not depend upon the correctness of that overarching vision of causation. All that I require is that the meaning postulate I associate with causal structure be a reasonable one.

Despite the pluralistic spirit expressed in the foregoing paragraph, it is important to stress that not any old thing can be an acceptable meaning postulate. For instance, it would be absurd to say that causation is that which is located in the top drawer of my desk. Absurd proposals like this one would clearly be disqualified by the requirement that a meaning postulate indicate something that is generally regarded as both evidence for and a consequence of causation. But some things that are conceptually linked to causation also fail this criterion. Suppose one proposed this as a meaning postulate: "Causation is that which is necessary for moral responsibility." That there is some conceptual link between moral responsibility and causation seems clear enough. In many cases, one can be morally responsible for something only if one has some influence on it. However, moral responsibility is not something that could serve as *evidence* for causation. Evidence for causation is something that you can actively search for or produce in order to decide whether a causal relationship obtains. If you want to know whether A causes B, you might do an experiment in which you manipulate A and check to see if B varies concomitantly. Or you might collect statistical data to see if A and B are correlated even when potential common causes are statistically controlled for. But there is no analogous way to use moral responsibility as a basis for testing causal claims. The same point would go for the suggestion that causation is that which underlies explanation. Consequently, not everything that is conceptually linked to causation can serve as a good meaning

3.3 CAUSAL STRUCTURE

A lively body of work on the problem of causal inference from statistical data uses directed graphs to represent causal structures (cf. Glymour and Cooper 1999; Spirtes, Glymour, and Scheines 2000; Pearl 2000; Neopolitan 2004). For example, consider Figure 3.1.

As in section 2.1, the nodes of the graph correspond to variables and an arrow from one node to another indicates the relationship of direct causation. For instance, Y might represent whether or not a particular power strip is switched to the "on" position, while X and Z each indicate whether or not an electrical appliance plugged into the power strip is on. Using directed graphs to represent causal structures has several advantages for theories of causal inference, the most significant of which is that it enables one to draw upon mathematical results which facilitate computationally tractable methods of deriving predictions about probabilistic independence and conditional independence from alternative causal hypotheses.[5] Directed graphs in conjunction with probability distributions are sometimes referred to as *Bayesian networks*, or *Bayes nets* for short.[6] For convenience, I shall adopt the label *causal Bayes nets* to refer to the approach to causal inference just briefly described.

Causal structures, then, are what directed graphs are intended to represent in the causal Bayes nets literature. But that does not tell us very much about what causal structures are; after all, directed graphs like that in Figure 3.1 can just as easily be used to represent mere correlations. And of course, things other than directed graphs—such as systems of equations and wiring diagrams—can also be used to represent causal structures. What is it, then, that these diverse modes of representation depict? Introductions to treatises on the topic typically emphasize the importance of causal inference for accurately predicting the consequences of public policy decisions (cf. Glymour and Cooper 1999, xi–xii; Pearl 2000, 337; Spirtes, Glymour, and Scheines 2000, xiii–xiv). In addition, significant effort is dedicated to inquiring how knowledge of causal structure, in varying degrees of precision, can serve as the basis of predicting consequences of interventions (cf. Spirtes, Glymour, and Scheines 2000, chap. 7). Thus, causal structures provide information concerning the results of interventions. An additional role is also attributed to causal structures: causal structures are said to "generate" probability distributions (cf. Glymour 1997, 206; Spirtes, Glymour, and Scheines 2000, 29).

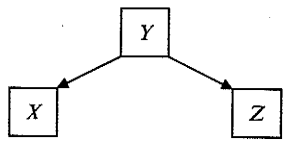


Figure 3.1 A directed graph

Pulling these two strands together, we have the following meaning postulate:

> (CS) *Causal structure* is that which generates probability distributions and indicates how these distributions will change given interventions.

A good understanding of (CS) is evidently dependent on some explication of interventions and of what it is to "generate" a probability distribution. Since the notion of an ideal intervention was explained in section 2.1, let us consider the second of these two questions.

For our purposes, the concern is with physical probability rather than probabilities interpreted as personal degrees of belief or confidence. Although the concept of physical probability is nearly as disputed as that of causation, I think that it is clear enough what sort of phenomena such probabilities usefully represent, namely, processes whose outcomes exhibit what John Venn described as a combination of "individual irregularity with aggregate regularity" (1962, 4). For example, consider the simple case of a flipped coin.

> So long as we confine our observation to a few throws at a time, the series seems to be simply chaotic. But when we consider the result of a long succession we find a marked distinction; a kind of order begins gradually to emerge, and at last assumes a distinct and striking aspect. We find in this case that the heads and tails occur in about equal numbers, that similar repetitions of different faces do also, and so on. In a word, notwithstanding the individual disorder, an aggregate order begins to prevail. (Venn 1962, 5)

As Venn observed, this type of behavior is found in many other circumstances: "Fires, shipwrecks, yields of harvest, births, marriages, suicides; it seems scarcely to matter what feature we single out for observation" (1962, 6).

For our concerns, it is unimportant whether one wishes to define probability as the aggregate or macro pattern itself (as frequency interpretations do), or as the causal tendencies underlying that aggregate pattern (as propensity interpretations do). Probabilities are useful for representing, or modeling, any phenomenon that displays a combination of individual irregularity and aggregate regularity. A process can be said to generate a probability distribution, then, just in case it gives rise to an aggregate pattern of this sort. This criterion is, admittedly, somewhat vague, but it will suffice for the present purposes.

Things that generate probability distributions, then, must exhibit behavior possessing the combination of individual disorder and aggregate regularity described by Venn. I maintain that these properties are possessed by mechanisms that are impinged on by disturbances that are, from the perspective of human knowledge, largely random. Moreover, mechanisms often provide information about the effects of interventions. Consequently, mechanisms are promising candidates for causal structure.

Let us consider this thought in more detail with regard to a pair of cases: molecular biology and social science.

3.4 CAUSAL STRUCTURE IN MOLECULAR BIOLOGY

Given a meaning postulate, the next stage of an empirical analysis is a contingent hypothesis, which specifies a class of entities whose extension, in a particular domain, is exactly that of the meaning postulate. In this section, I argue that in molecular biology, causal structure coincides with mechanisms, yielding the following empirical analysis:

- *Meaning Postulate* (CS): *Causal structure* is that which generates probability distributions and indicates how these distributions change under interventions.
- *Contingent Hypothesis*: In molecular biology, *mechanisms* are what generate probability distributions and indicate how these distributions change under interventions.
- *A Posteriori Identity*: In molecular biology, mechanisms *are* causal structure.

In this section, I argue in favor of the above contingent hypothesis. As explained in earlier sections of the chapter, empirical analyses rely upon established scientific theories of the relevant domain. In this case, evolutionary biology plays an important role in motivating the claim that mechanisms in molecular biology provide information about the consequences of interventions by providing a general reason to expect that such mechanisms are modular.

3.4.1 What's a Mechanism?

Mechanisms, in a very literal sense of the term, are paradigmatic examples of causal structures. For example, in Nancy Cartwright's words:

> The car engine is a good case of a stable causal structure that can be expected to give rise to a probability distribution over the events of the cooperating causal processes that make it up. That is why it can make sense to ask about the conditional expectation of the acceleration given a certain level of the throttle. (1995a, 72)

Given that several authors have proposed that mechanisms play an important role in the life sciences (cf. Bechtel and Richardson 1993; Glennan 1996; Machamer, Darden, and Craver 2000), they are a natural place to turn for an empirical analysis of causal structure in biology. However, this must be done with some care, since the application of the word "mechanism" in distinct domains might reflect only a superficial similarity of subject matter. Thus, it is important to examine just what sorts of things biological mechanisms are and why they should be thought to fulfill the roles ascribed to causal structure.

Mechanisms are generally understood as consisting of interacting components that generate a causal regularity between some specified beginning and end points. For example, according to a definition proposed by Peter Machamer, Lindley Darden, and Carl Craver, "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000, 3). This general characterization is appropriate for literal examples of mechanisms, such as the car engine, and is reasonable with regard to things referred to by the term "mechanism" in biological science. Consider, for example, the mechanism involved in protein synthesis, in which the series of nucleotide bases in strands of DNA influences the chemical structure of proteins produced within cells. Nearly any introductory biology textbook describes this mechanism roughly as follows. First, a strand of DNA unwinds and the adjoining nucleotide bases separate. The next step is the transcription of the unwound DNA by messenger RNA (mRNA), the order of the bases of the mRNA being determined by the order of the complementary nucleotide bases in the DNA strand. Finally, the strand of mRNA serves as a template for transfer RNA (tRNA), which assembles a string of amino acids into a protein. In this case, the interworking parts give rise to more readily observed regularities, such as correlations between genes and specific traits. Some things referred to by the term "mechanism" may not involve a regular series of changes. For example, the term "mechanism" is sometimes used to refer to a unique chain of events leading to a particular effect. However, since this book is concerned with extrapolating causal generalizations, I will use the term "mechanism" to refer to regularly operating causal relationships rather than idiosyncratic and unique chains of events. Consequently, I will restrict the term "mechanism" to processes that satisfy the "regular changes" clause of the Machamer-Darden-Craver definition.

Other related definitions of mechanisms exist. For example, Stuart Glennan proposes a definition that is similar to Machamer, Darden, and Craver's except that it requires that the interactions among the components of the mechanism be governed by "direct causal laws" (1996, 52). The reference to laws in this definition is problematic, since it is debatable whether there are genuine laws of nature in biology and social science, where the term "mechanism" is often used. Consequently, in a subsequent revised account of mechanisms, Glennan replaces "direct causal laws" with "direct, invariant, change relating generalizations" (2002, S344). The notion of an invariant generalization is borrowed from James Woodward (2000, 2003). An invariant generalization is one that is invariant under some range of ideal interventions on the allegedly explanatory variable. For example, the generalization that barometer readings and storms are correlated is not invariant under ideal interventions on the barometer readings (as explained in section 2.1). Hence, the barometer readings do not cause or explain storms, according to Woodward's the-

cancer would be invariant under ideal interventions that target smoking. James Tabery (2004) argues that there is an important difference between Woodward's conception of causation and the notion of "productivity" invoked in the definition proposed by Machamer, Darden, and Craver. The thought is that while invariant generalizations merely point to ways in which changes brought about by an intervention lead to specific changes someplace else, productivity pertains as well to cases in which new entities are constructed (2004, 8–9). However, the "changes" covered by Woodward's account of causation should be understood to include constructing a new product out of disparate parts. For example, imagine a cellular process that generates a particular enzyme. Let E be a variable that indicates whether or not this enzyme has or has not been produced on given occasions. Then there may be invariant generalizations relating E to other variables that represent, say, the presence of necessary components in the cell or the transcription of a particular gene. If there is a real difference between Glennan's definition and that proposed by Machamer, Darden, and Craver, I think it is only that Glennan provides more detail about his preferred interpretation of causation.

Cartwright's *nomological machine* is another mechanism concept. Cartwright defines a nomological machine as "a fixed (enough) arrangement of components, or factors, with stable (enough) capacities that in the right sort of stable (enough) environment will, with repeated operation, give rise to the kind of regular behaviour that we represent in our scientific laws" (1999, 50). Like the definitions of mechanism considered above, Cartwright's nomological machine consists of interacting components that generate causal regularities. The concept of a nomological machine is distinctive only insofar as it is founded on Cartwright's concept of a capacity. A capacity is a stable causal power that exerts its characteristic influence in a broad range of contexts (Cartwright 1989, Chapter 4). The pure effects of a capacity can be observed only in special experimental circumstances in which all other causes have been eliminated, but the capacity nevertheless makes its contribution to the effect even when other causes are present. Since Cartwright regards physical laws merely as descriptions of the behavior of a capacity in the idealized situation in which no other forces are acting, she regards capacities as ontologically more basic or fundamental than laws of nature. Cartwright also argues that interpreting causal relationships by reference to capacities is essential for understanding how it is possible to extrapolate causal claims from one context to another (1989, 163). I argue in Chapter 5 that capacities do not in fact have this special virtue. But for the moment, let us sum up the above survey of mechanism concepts.

All of the definitions canvassed above characterize mechanisms as consisting of sets of interacting components that generate a regular series of causal interactions. To the extent that they disagree, it is with regard to how to interpret causation. For example, Glennan's original definition (1996) characterized causation by reference to "direct causal laws," while

Cartwright prefers capacities. Fortunately, pursing an empirical analysis of causal structure does not require deciding whether laws or causal powers are more fundamental or insisting that there is one correct way to interpret causation. Instead, it requires an argument that mechanisms generate probability distributions and provide information about how those distributions change under interventions. Given this, I will adopt the Machamer–Darden–Craver definition, since it is the least specific about causation, laws, and their relation to one another. The question, then, is whether mechanisms, so defined, are causal structures. I consider this question first with regard to molecular biology and then for social science.

3.4.2 Mechanisms, Modularity, and Evolvability

There is good reason to think that if there is such a thing as causal structure in molecular biology, it would have to be mechanisms. First, note what might be called the working assumption of molecular biology: all causal relationships in living organisms are mediated by molecular processes. This working assumption rests on the attractiveness of physicalism as a general ontological principle and on the success of molecular biology as a research program. Thus, if mechanisms are not causal structures in molecular biology, it is hard to see what could be. However, this conclusion is only half of the argument. It is also necessary to show that mechanisms in molecular biology do in fact perform the functions required of causal structure.

Since causal structure is that which generates probability distributions and provides information about how those distributions change given interventions, there are two parts to this argument. Let us begin with the requirement that causal structure generate probability distributions. Is this something that mechanisms in molecular biology do? Recall the features that Venn judged to be characteristic of phenomena to which the concept of probability can be usefully applied: individual disorder combined with aggregate regularity. It is obvious that mechanisms in the sense of the Machamer–Darden–Craver definition will tend to generate large sample regularities, given the requirement that mechanisms "are productive of regular changes" from the beginning and end stages of the process. Moreover, biological mechanisms are invariably subject to an array of disturbing influences, many of which are not well understood. Thus, from the perspective of human knowledge, individual cases of the operation of a given mechanism in molecular biology will inevitably display a certain amount of random variation, which is an example of the "individual disorder" that Venn described. Notice that the same sort of situation is found in the case of human-constructed machines, which are often given as paradigm examples of causal structure. They produce regular changes, yet are impinged upon by a variety of disturbing influences that often cannot be known with any exactitude. Consequently, we

should display the aggregate regularity and individual disorder that Venn cited as the characteristic features of probabilistic phenomena. Of course, these aggregate patterns may themselves change in the course of evolution, but this simply illustrates the familiar point that probability distributions themselves can change over time (cf. Venn 1962, 14–17). This point is illustrated by such social statistics as the marriage rate or average life span. Indeed, it is exemplified by Cartwright's case of the car engine; the probability of a breakdown increases as the engine ages.

However, since knowledge of causal structure also provides information about the consequences of interventions, an account of why mechanisms in molecular biology should be thought to generate probability distributions is only half of the story. It is necessary to argue that mechanisms in molecular biology generally provide information about the results of interventions. On the face of it, it is quite plausible that this is the case. Indeed, this presumption that knowledge of mechanisms can indicate the consequences of various types of interventions is often the reason for trying to discover them. But is there some general feature of biological mechanisms that justifies this presupposition? One answer to this question has been suggested by Woodward (2002a, S374–76), who maintains that mechanisms are *modular* in the sense that it is possible to intervene to change a feature of one component while leaving the generalizations that govern the others unaltered. This idea is reflected in the manner in which interventions are represented in directed graphs. Consider again the case of the two appliances plugged into the same power strip, represented by the graph in Figure 3.1. Recall that an ideal intervention takes complete control of the variable it targets (say, X), so as to eliminate all other influences that otherwise affect it. Such an intervention, as we saw in section 2.1, would be represented as shown in Figure 3.2.

Of course, many real-life interventions are not ideal. In our example, switching on one of the appliances would not be an ideal intervention, since it does not sever the influence of the state of the power strip. Such an intervention might be represented as shown in Figure 3.3:

The important point with regard to modularity in figures 3.2 and 3.3 is that besides possibly eliminating or weakening the influence of Y upon X, the intervention leaves all other causal relationships unaltered. For example, modularity would be violated if the intervention eliminated the influence of Y upon Z or created a causal chain from X to Z. The interest in modularity stems from the fact that it facilitates predicting the

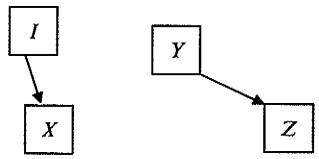


Figure 3.2 An ideal intervention

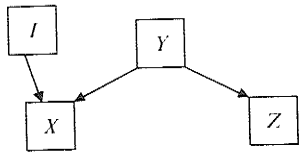


Figure 3.3 A nonideal intervention

consequences of interventions, since except for the elimination of influences upon the targeted variable, the causal structure operates as before. Thus, the persisting causal structure can be used to trace out the intervention's consequences. Given Woodward's proposal, the question is whether there is reason to believe that biological mechanisms are usually modular. Some classic examples of biological mechanisms do exhibit modularity. For example, it is possible to alter the sequence of nucleotide bases in a functional segment of DNA while the other components in the mechanism of protein synthesis continue to function as before. But is modularity only an adventitious feature of some restricted class of biological mechanisms, or is there some reason for supposing that it obtains in general?

One way to argue that modularity is likely to be a commonly occurring feature of biological mechanisms is to maintain that modularity is favored by natural selection. Herbert Simon (1962) was one of the first to suggest a general explanation of how modularity is adaptively beneficial in environments in which disruption or interference is common. The proposal can be illustrated with a modified version of one of Simon's best-known examples, the parable of the expert watchmakers Hora and Tempus (1962, 470).[7] Hora constructs her watches by building independently changeable modules that can be assembled into the final product. In contrast, Tempus constructs holistic watches in which no part can be modified independently of any of the others. Hora's modular production method gives her an advantage over Tempus as their ever more popular watches are used in new circumstances. For instance, mountain climbers find that the watches of Hora and Tempus fail to operate properly at high altitudes. Hora is able to trace the problem to a specific module, and through trial and error she develops a new version that operates properly under high-altitude conditions. In contrast, Tempus must redesign an entirely new high-altitude watch, which means searching for a solution through the space of possible watches, which is far vaster than the space of possible modifications of a specific component. By the time Tempus has finished his holistic high-altitude chronometer, Hora has already cornered the mountain climber watch market, as well as that for scuba divers, mariners, pilots, runners, and several other specialty niches. The moral of the parable, then, is that modularity facilitates finding quick solutions to new problems, which is essential for adapting to changing

The theme of this parable is nicely illustrated by the HIV replication mechanism that will be discussed in detail in the next chapter. HIV is notorious for its ability to evolve resistance to drugs designed to block its replication. Typically, such drugs interfere with one stage of the replication mechanism, for example, by binding to and disabling an enzyme required for a step in the process. In this case, a mutation in the viral genome can result in a slightly modified version of the enzyme to which the therapeutic compound no longer binds. Given that the other components of the mechanism continue to function as before, HIV has successfully evolved resistance; but if the change to the enzyme resulted in cascading alterations to the other components, it is likely that the mutant strain would no longer be viable. Thus, the HIV replication mechanism is analogous to Hora's production method: since it is modular, alterations to one component to do not compromise the functionality of the others. Consequently, evolving resistance to a single drug requires altering only one component of the replication mechanism, and hence searching through a smaller space of possibilities. In contrast, if the HIV replication mechanism were holistic like Tempus's watches, evolving resistance to the therapeutic compound would require rebuilding the mechanism from scratch, and hence searching for a solution in the space of all possible HIV replication mechanisms. Thus, modularity is an important part of what enables HIV to quickly evolve resistance.

These examples suggest that modularity enhances fitness by promoting adaptability to changing environments. Moreover, environmental perturbations of various kinds—new predators, changes in supply of resources, and so on—are a pervasive fact of life. Hence, evolutionary theory suggests a basis for expecting that modularity is a typical characteristic of biological mechanisms. In fact, the importance of modularity to adaptability is a familiar point in evolutionary biology (cf. Wagner and Altenberg 1996). There is a growing body of theoretical work that attempts to clarify the general mechanisms whereby natural selection could give rise to modularity (cf. Ancel and Fontana 2000; Lipson et al. 2002; Kvasnicka and Pospichal 2002; Kashtan and Alon 2005). This work supports the intuition that natural selection favors modularity in changing environments, but with some refinements. For example, one recent study suggests that although not all varying environments lead to modularity, modularity is favored in environments with "modularly varying goals" (Kashtan and Alon 2005, 13777). Goals vary modularly when new goals share subproblems with preceding goals (ibid., 13775). The HIV example illustrates this concept. At first, the goal of the enzyme is to achieve a particular function, say, to reverse transcribe viral RNA to DNA. After the start of the drug treatment, the enzyme must still perform its original function while also avoiding being bound to the therapeutic compound. Hence, reverse transcribing the viral RNA to DNA is a subproblem shared by the first and second goals. The situation in the watchmaker parable is similar. In redesigning the malfunctioning module, Hora

must preserve its original function while avoiding the disruption that occurs at high altitudes. Modularly varying goals might drop as well as add subproblems. For instance, consider a population of fish that has colonized a network of underground pools: the fish no longer need to see, but they still need to swim.

There is also a growing number of empirical studies that examine the role of modularity in the evolution of particular lineages (cf. Beldade et al. 2002; Chipman 2002; Mabee et al. 2002; Friedman and Williams 2003; Emlen et al. 2005; Fraser 2005).[8] These studies provide fascinating concrete examples of the ways in which modularity can be manifested in living beings. For example, one study documents how threshold mechanisms allow for developmental modularity in the evolution of beetle horns (Emlen et al. 2005). Empirical studies can also test hypotheses about the relationship between modularity and evolvability. For instance, mixing and matching modules, sometimes called "compositional evolution," may often be a more efficient means of finding a solution to a problem than randomly rearranging basic elements (Watson and Pollack 2005, 456). By analogy, one is more likely to produce a sentence by randomly combining clauses and phrases than by randomly combining letters and spaces. An additional potential advantage of compositional evolution, in contrast to gradual accumulation of slight variations, is that it can avoid suboptimal local maxima traps, since a rearrangement of modules constitutes a jump to a nonadjacent point in the fitness landscape (Kashtan and Alon 2005, 13777). And in fact a recent study finds support for compositional evolution with regard to protein modules in yeast (Fraser 2005). In the HIV example discussed above, compositional evolution would suggest that the resistant variant resulted from rearranging proteins that compose the enzyme rather than from shuffling the individual amino acids that make up the proteins.

Mechanisms that are modular in the sense of these biological discussions are ipso facto a useful basis for predicting the consequences of interventions. Although several modularity concepts can be found in biology (Schlosser and Wagner 2004), the following is a fairly standard, rough definition that is appropriate for the present context:

> A modular representation of two character complexes C1 and C2 is given if pleiotropic effects of the genes fall mainly among members of the same character complex, and are less frequent between members of different complexes. (Wagner and Altenberg 1996, 971)

According to this definition, modularity states that the multiple effects of genes tend to focus on discrete trait complexes. This definition makes the connection between modularity and manipulability straightforward. For if modularity in the sense just defined obtains, it is possible, by means of appropriate alternations to the genome, to intervene to alter one component of the mechanism without significantly disturbing the others. Thus, knowledge of modular mechanisms would provide information about the

consequences of interventions. Of course, it would be a mistake to take the above as a *general definition* of modularity. Rather, it is a rough specification of the physical basis of modularity in molecular biology—in effect, an empirical analysis of modularity in that context. An empirical analysis of modularity in social science, for instance, would have to be something rather different.

In sum, given the meaning postulate that causal structure is that which generates probability distributions and indicates how such distributions change given interventions, evolutionary theory plays a central role in an empirical analysis of causal structure in molecular biology. Evolutionary theory can be invoked to support the claim that in the context of molecular biology, mechanisms can be identified with causal structure, since it provides an account of why it should be expected that biological mechanisms are typically modular. Modularity, meanwhile, was linked to the ability to predict the consequences of interventions. Of course, since empirical analysis depends on current scientific theory, it is inherently tentative. New scientific developments might result in significant revisions to the theory, and these developments might have implications for the empirical analysis. The evolution of modularity in biological systems is a young and thriving research area, which means that we should expect surprises yet to come.

3.5 CAUSAL STRUCTURE IN SOCIAL SCIENCE

In this section, I consider the possibility that an empirical analysis identifying causal structure with mechanisms in molecular biology on the basis of evolutionary theory could work similarly in social science. On its face, the argument for the adaptive benefits of modularity in variable environments seems entirely general, and hence applicable to cultural as well as to biological evolution. However, the details of these proposals are at present far less developed in social science than in biological science. In addition, one common argument against the possibility of laws of social science can be interpreted as an attempt to show that social mechanisms will often respond in nonmodular ways to interventions. Thus, I conclude that although it is likely that the evolutionary account of modularity described above can be applied to some social mechanisms, the extent to which this is so is even more of an open question than in the case of mechanisms in molecular biology.

3.5.1 What's a Social Mechanism?

In order to consider whether social mechanisms are likely to be modular, some clarification of "social mechanism" is called for. Earlier, mechanisms in general were roughly characterized as sets of entities and activities organized so as to produce a regular series of changes from a beginning state to an ending one. Social mechanisms in particular are usually thought of as complexes of interactions among agents that

underlie and account for macrosocial regularities (cf. Little 1991, 13; Stinchcombe 1991, 367; Schelling 1998, 33; Gambetta 1998, 102). The paradigm example of an agent is an individual person, but coordinated groups of individuals motivated by common objectives—such as a corporation, a government bureau, or a charitable organization—may also be treated as agents for certain purposes (cf. Mayntz 2004, 248). Social mechanisms are sometimes tied to the assumption that the agents comprising them are rational, say in the sense of being utility maximizers. For instance, Tyler Cowen writes, "I interpret social mechanisms . . . as rational-choice accounts of how a specified combination of preferences and constraints can give rise to more complex social outcomes" (1998, 125). I shall not adopt this perspective, and hypotheses about social mechanisms will not be restricted to rational-choice models.

Social mechanisms typically involve reference to some categorization of agents into relevantly similar groups defined by a salient position their members occupy vis-à-vis others in the society (cf. Hernes 1998; Little 1998, 17; Mayntz 2004, 250–52). In the description of the mechanism, the relevant behavior of an agent is often assumed to be a function of the group into which he or she is classified. For example, consider the anthropologist Bronislaw Malinowski's (1935) account of how having more wives was a cause of increased wealth among Trobriand chiefs. Among the Trobrianders, men were required to make substantial annual contributions of yams to the households of their married sisters. Hence, the more wives a man had, the more yams he would receive. Yams were the primary form of wealth in Trobriand society, and served to finance such chiefly endeavors as canoe building and warfare. Although individuals play a prominent role in this account, they do so as representatives of social categories: brothers-in-law, wives, and chiefs. The categorization of component entities into functionally defined types is not unique to social mechanisms. Biological mechanisms (e.g., that of HIV replication) are often described using such terms as "enzyme" and "co-receptor." The terms "enzyme" and "co-receptor" resemble "chief" and "brother-in-law" in virtue of being functional: all of these terms provide some information about what role the designated thing plays in the larger system of which it is a part. In sum, social mechanisms can be characterized as follows. Social mechanisms are complexes of interacting agents—usually classified into specific social categories—that produce regularities among macrolevel variables.

This characterization of a social mechanism can be illustrated by another, better-known example. Consider Thomas Schelling's bounded-neighborhood model, which is intended to account for persistent patterns of segregated housing in spite of increased racial tolerance (Schelling 1978, 155–66). In this model, the residents of a given neighborhood are divided into two mutually exclusive groups (e.g., black and white). Each individual prefers to remain in the neighborhood, provided that

threshold, which may vary from person to person. Meanwhile, there is a set of individuals outside the neighborhood who may choose to move in if the proportions are to their liking. Clearly, this model divides individuals into groups with which characteristic preferences and subsequent behavioral patterns are associated, and by these means accounts for macroregularities.

On the face of it, it might seem that the empirical analysis of causal structure given in section 3.4 easily transfers to social science. As in the case of molecular biology, it is difficult to see what could constitute causal structure in social science if not social mechanisms. Moreover, it is plausible that social mechanisms often produce stable patterns, and hence generate probability distributions. Finally, just as in the case of biology, it seems that modularity is a feature that contributes to the adaptability of social systems. Indeed, the parable of Hora and Tempus illustrates the advantages of modularity for technology and is analogous to such historical cases as the IBM PC versus the Apple Macintosh, and General Motors versus Henry Ford (cf. Langlois 2002, 23–33). However, it is unclear how far the evolutionary argument for the prevalence of modularity carries over to the social realm.

3.5.2 Modularity and Social Mechanisms

Let us consider how the evolutionary argument for modularity described in section 3.4.2 might work with regard to social phenomena. As a first stab, consider the following suggestion. Modular social mechanisms contribute to the adaptability of the social groups containing them. Such groups would be able to adapt more quickly to modularly varying environments by altering one module while leaving the others the same or by rearranging modules. And, as in biology, modularly varying environments are a pervasive fact of social life: human social groups often need to develop the capacity to solve new problems while retaining most of their prior problem-solving abilities. Thus, groups possessing modular mechanisms would be more likely to survive and produce "offspring" in the form of offshoot or copycat groups or organizations. However, there is reason for skepticism about this scenario.

The unit of selection in the scenario just described is the social group, and one important type of social group is the organization. In fact, there is a social science research program inspired by evolutionary biology in which the units of selection are organizations, namely, organizational ecology. Organizational ecology attempts to explain characteristics of various types of organizations—businesses, labor unions, advocacy groups, churches, and so on—in distinct contexts on the basis of differential mortality and founding rates (cf. Hannan and Freeman 1989; Aldrich 1999, 43–48). For example, one important thread in this literature examines the distinct environments to which generalist and specialist organizations are best suited, for instance, inquiring into the conditions in which consolidation among generalist organizations creates resource

opportunities for specialists (cf. Carroll and Swaminathan 2000). Unfortunately, the scenario sketched in the foregoing paragraph contradicts one of the basic premises of organizational ecology: the structural inertia of organizations (Hannan and Freeman 1989, 70; Aldrich 1999, 45). According to this principle, the rate of change of an organization's structure is typically much slower than the rate of change in the environment. This premise is important for a model in which Darwinian selection is the driving force. Changes in populations of organizations result primarily from old organizations disbanding and being replaced by new ones that are better suited to the new environment rather than from individual organizations adapting themselves to new situations. There are a number of reasons why organizations would be expected to exhibit structural inertia (Hannan and Freeman 1989, 67–69). For example, restructuring often shifts resources away from a segment of the organization, and hence is likely to be resisted by those members who would be disadvantaged. Moreover, there is some empirical evidence in support of structural inertia (Aldrich 1999, 168). Thus, the proposal that highly modular, and therefore quickly changeable, organizations are favored by social selection processes is problematic.

Let us try a different approach. Modularity of social mechanisms need not entail that individual organizations be quick to adapt to changing circumstances. That point can be appreciated through a consideration of modular mechanisms in molecular biology. In that case, modularity is a matter of how the genome maps onto system components, not a claim that *individual* organisms can quickly adapt to new environments. The adaptation that modularity engenders, occurs across generations, not in the life history of a single organism. Thus, perhaps things work similarly in the social world. Consider two general ways in which this might happen.

First, consider social mechanisms that are internal to organizations. These mechanisms might include such things as a social hierarchy or an established production procedure. In this case, the argument would be that modularity facilitates evolvability because it allows mechanisms to be modified one component at a time or for solutions to new social problems to be found by rearranging mechanism components. This scenario is consistent with structural inertia, since the altered versions of the mechanism might occur in newly founded organizations rather than in transformed versions of older ones. In this scenario, nonmodular social mechanisms would be likely to go extinct in modularly varying environments, while the varied descendants of modular mechanisms would spread throughout the population of organizations. The plausibility of this scenario is enhanced by the wide prevalence of certain types of modular structures found in organizations and social groups in general, particularly hierarchies. For example, consider the hierarchical structure of a university: the university is divided into colleges or schools, which are in turn divided into departments or units. This structure is modular,

philosophy department) while leaving other units as they were before. Likewise, although it would be difficult for an established university to, say, eliminate a number of existing departments or to restructure its colleges, a newly founded university might readily make such changes.

A second scenario concerns social mechanisms that are not internal to specific organizations, but instead are features of the broader social context in which organizations as well as individuals are embedded and interact. Forms of economic interaction, such as a market, are examples of social mechanisms of this kind. Again, the hypothesis would be that such mechanisms, if modular, are more adaptable to changing environments. As a result, such mechanisms would be expected to proliferate more widely than their nonmodular counterparts. An economic system based upon property rights and market exchange is arguably a modular mechanism, since it allows owners wide leeway to modify their properties or enterprises independently of others (cf. Langlois 2002, 26–27). Such a system also allows for rearrangement of modules in the form of consolidation or increasing specialization of industries. A more specific example is the contrast between traditional and Silicon Valley models of research and development (Aoki and Takizawa 2002). In the traditional model, R&D is carried out in an integrated manner within a particular firm, which organizes and directs coordinated R&D projects for specific goals. In this model, it is important that each of the various design teams knows what the others are doing, so that their results can be assimilated into the final product. Clearly, communication among design teams becomes increasingly cumbersome with the increasing complexity of the task of each. In the Silicon Valley model, by contrast, the product system is divided into modules developed by separate firms, often start-ups funded by venture capitalists. The Silicon Valley model requires standardized interfaces between modules, so that improvements to the overall product system result primarily from independent improvements in the various components (Aoki and Takizawa 2002, 770–71). The advantage of the Silicon Valley model is that it avoids the onerous communication among design teams required by the traditional model, thereby facilitating quicker solutions to new problems. The Silicon Valley model, then, is an example of a modular mechanism that structures the interactions of a collection of organizations. But there is nothing in this scenario to require that individual organizations be highly adaptable.

The two scenarios described above illustrate ways in which the hypothesis about the advantages of modularity with regard to evolvability might be extended to social mechanisms. But the quantity of both theoretical and empirical research on these questions in social science is minuscule in comparison to the body of work on modularity and evolvability in biology. Robert Boyd and Peter Richerson (Richerson and Boyd 2005; Boyd and Richerson 2005) are the only authors I know of who have offered anything like a detailed evolutionary explanation of modularity in social science. Boyd and Richerson argue against the image of culture

as a tightly integrated, holistic system (Richerson and Boyd 2005, 91–93), and they hypothesize that culture evolved as an adaptation to rapidly changing climates in the Pleistocene (ibid., 131–39). They develop models that illustrate how the cumulative social learning indicative of culture can be favored by natural selection in changing environments (Boyd and Richerson 2005, pt. I). The main theme of this account is that culture enhances adaptability by facilitating quick, though not necessarily optimal, solutions to new problems. Hence, Boyd and Richerson's hypothesis is very similar to the evolutionary account of modularity described in section 3.4.2. Nevertheless, the focus of Boyd and Richerson's work is explaining the origin of culture rather than modularity per se, and it is unclear to what extent their proposals could be developed to support the claim that specific types of social mechanisms are modular.

In the remainder of this section, I consider some possible reasons for thinking that social mechanisms may often be nonmodular. The first concern is based on the point that modularity is adaptively beneficial only in changing environments. Consequently, nonmodular designs may be preferable to modular ones in environments that exhibit a high degree of stability over time. Thus, there would appear to be no particular reason to expect modular social mechanisms in social contexts that have persisted without much change for a significant period. Richard Langlois suggests that certain nonmodular features of medieval European social structures were well suited to the stable social environment of this period, but eventually disappeared in the face of changing circumstances (2002, 28–29). Of course, the analogous point holds with regard to biology as well. Thus, the question here is to what extent past social and biological environments have been modularly variable rather than stable or simply chaotic. The next concern, however, is more specifically focused on characteristic features of human society.

A common challenge for social policy is that changes in one feature of a society may produce unpredictable changes elsewhere in the system, thus making it extremely difficult to anticipate the consequences of the policy intervention. One source of this difficulty is that participants in the system who are not directly targeted by the policy intervention may nevertheless be aware of it, and may perceive opportunities to advance their interests by modifying their practices in response to it. Indeed, the complex interrelation between social structures and awareness of those structures by members of the society is a common basis for arguments against the possibility of laws of social science (cf. Searle 1984; Taylor 1971). Although such arguments rarely use the term "modularity," the modularity of social mechanisms is precisely what they aim to call into question. For if the objection is correct, it will typically not be possible to change one component of a social mechanism without producing unpredictable changes in the others.

This objection to the modularity of social mechanisms will be discussed

that would be relevant to any response to it. Whether a mechanism is modular with regard to an intervention depends on the intervention itself and on the manner in which the causal system is represented. For a given mechanism, some interventions may be modular while others are not. In Chapter 8, I call interventions that affect mechanisms in nonmodular ways *structure-altering*. The second point is that even if an intervention is structure-altering with regard to a mechanism, it might not be such with regard to other, more fundamental mechanisms that can explain why and how the first was altered. Thus, one natural response to the objection described in the foregoing paragraph is that the unintended effects of the policy intervention could be explained, and perhaps even anticipated, by individual-level mechanisms. For example, a rational choice model might explain why an intervention that inadvertently creates new incentives leads to systematic but unintended changes of behavior. The thought that more fundamental, modular mechanisms can be described at finer-grained levels of description is an underlying motivation of the mechanisms approach to extrapolation. It is also the central theme of Chapter 7, which discusses the relationship between mechanisms-based extrapolation and reductionism.

3.6 CONCLUSION

This chapter began with the question of the relationship between mechanisms and the probabilistic causal concepts elaborated in Chapter 2, and it proposed the first part of an answer to this question. To the extent possible, mechanisms are to be identified with causal structure on the basis of domain-specific empirical analyses. Since causal structure is that which generates probability distributions and provides information about how they change under interventions, this identification is a basis for linking mechanisms to probabilistic causal concepts. An important part of these empirical analyses consists of providing some general reason to think that mechanisms are modular, and evolutionary theory suggests a means of doing just this. However, this evolutionary argument is, at present, on firmer ground in molecular biology than in social science.

Yet the identification of mechanisms with causal structure alone indicates only that there is *some* connection between mechanisms and probabilistic causal concepts such as causal effect and positive causal relevance. It provides no indication of what the nature of that relationship is. Chapter 4 discusses a proposition, which I call the *disruption principle*, which says something specific about the link between probability and mechanisms identified with causal structure.