# 4

# The Disruption Principle

*All of the known mechanisms by which HIV impairs the human immune system depend on HIV reproduction. Therefore, the development of anti-HIV replication drugs would appear to be a positive first step in controlling HIV reproduction.*

—Gerald Stine, *AIDS Update 2000* (2000, 84)

Implicit in the above quotation is a commonsense idea that I think is widespread in much of biology and probably in many other areas of science as well: a causal effect is completely nullified when, and only when, every mechanism linking cause and effect is severed. This is a rough statement of what I call the *disruption principle*, which is the main topic of this chapter. The disruption principle specifies a relationship between mechanisms, identified with causal structure as proposed in Chapter 3, and probabilistic causal concepts such as causal effect and causal relevance. In virtue of making this connection, the principle plays an important role in the treatment of extrapolation developed in subsequent chapters. The purpose of this chapter is to articulate the disruption principle, illustrate it by reference to a scientific example, and explore the range of circumstances in which it can be reasonably presumed.

Since a concrete example greatly facilitates the presentation of the disruption principle, I begin with a brief description of the mechanism of HIV replication, and of some of the factors known to interfere with it. I then turn to a statement of the principle itself. Formulating the disruption principle in a precise way involves developing a graphical framework for representing factors that disrupt mechanisms, so this is somewhat complex. I then illustrate the disruption principle and the graphical framework in question by means of an example drawn from HIV research, namely, the discovery of a genetic mutation that confers substantial resistance to HIV infection. In the abstract, the problem illustrated by this example takes the following form. Suppose we know that a certain causal relationship holds between $X$ and $Y$ in the population P, for example, that $X$ is a positive causal factor for $Y$. We want to know if there is a subpopulation of P in which this effect is nullified. From the disruption principle it obviously follows that such a population must be one in which every mechanism from cause to effect is blocked. The challenge, then, lies in ascertaining whether such a subpopulation exists, given a lack of full knowledge of the total set of mechanisms, a state of incomplete knowledge



Given the important role of the disruption principle, it is worthwhile to consider what justification there is for assuming it. I show that, given the identification of causal structure and mechanisms argued for in Chapter 3, the disruption principle can be shown to follow from the conjunction of two more familiar principles concerning causality and probability: the principle of the common cause (PCC) and the faithfulness condition. This suggests that the disruption principle might be false—and hence the mechanisms approach to extrapolation, unreliable—when one or both of these principles do not obtain. That point is hardly trivial, since doubts have been raised concerning both the PCC and the faithfulness condition (cf. Sober 2001; Cartwright 1999). I argue that the PCC is on firm ground with respect to the types of cases that concern the disruption principle. The situation in the case of the faithfulness condition, in contrast, is more complex. I suggest that there is a strong motivation for the FC when studying heterogeneous populations, but that this justification collapses for exceedingly homogeneous populations, such as closely inbred strains of laboratory mice reared under uniform conditions. This result entails that special care should be taken to vary genetic or environmental background conditions in gene knockout experiments, a point which some researchers in this field have noted. That heterogeneity can be a virtue in scientific experiment is surprising in light of the common notion that the ideal experiment is one in which all factors except those subject to investigation are held constant.

## 4.1 HIV REPLICATION

According to the current standard, AIDS is diagnosed when a person's T-helper cell count drops below 200 per microliter of blood (cf. Stine 2000, 132; Kalichman 1998, 78–79).[1] Hence, the role of T-helper cells in the human immune system is a good place to begin in describing HIV replication.

The primary actors of the immune system are a collection of distinct types of white blood cells that identify and destroy antigens present in the body. For example, phagocytes eat bacteria and foreign or infected cells, while mast cells and eosinophils attack intruders too big for consumption (e.g., worms) by emitting poisonous chemicals in their vicinity (cf. Fan et al. 2000, 26–27). However, the cells of greatest concern for our purposes are lymphocytes. These come in two basic varieties, the B lymphocytes and the T lymphocytes. T lymphocytes themselves come in several varieties, the most important of which for our purposes are T-helpers and cytotoxic T-cells, or T-killers. The B lymphocytes identify which entities in the body are to be attacked by phagocytes, mast cells, and eosinophils (ibid., 32–37). B lymphocytes perform this function by producing antibodies, which are proteins that attach to particular sorts of intruding agents. Different B lymphocytes produce different antibodies, the specific antibody produced being determined by the result of a random rearrangement

of DNA within the cell. Once mature, a B lymphocyte will not replicate itself or emit its antibodies into the bloodstream unless two things happen. First, it must encounter an antigen to which its antibody attaches. Next, it must encounter a T-helper cell that also attaches to this antigen; when this happens, the T-helper chemically signals the B lymphocyte to release its antibodies and commence replicating. Hence, in the absence of T-helpers, the immune system is unable to identify which bodies are to be destroyed (e.g., by phagocytes) and which are not.

T-killer cells differ from B lymphocytes in that they directly attack and destroy antigens, yet in other respects the functioning of the two types of cells are very similar (cf. Fan et al. 2000, 42–47). Like B lymphocytes, different T-killer cells have a chemical affinity for different types of antigens. Moreover, a T-killer cell that has encountered an antigen to which it attaches will not replicate until instructed to do so by a T-helper cell that recognizes the same antigen. Thus, like the B lymphocytes, the T-killer cells can perform their role within the immune system only in the presence of T-helper cells. It can easily be understood, then, how a large-scale reduction in the number of T-helper cells would lead to catastrophic failure of the immune system and to opportunistic infections.

Given this background, we can proceed to a description of the mechanism by which HIV infects and destroys T-helper cells. It should be noted that T-helper cells are not the only cells of the human immune system that are infected by HIV. For example, a type of phagocyte, namely the macrophage, is also prone to HIV infection. Indeed, in the early and nonsymptomatic stages, HIV infection is restricted almost exclusively to macrophage-tropic (M-tropic) HIV, with widespread infection of T-helper cells occurring later in the progression of the disease (Zhu et al. 1993).[2] Macrophages will play a significant part in the discussion in the following section. We will see there that a few lucky individuals possess a genetic mutation that blocks HIV entry into macrophages, thereby conferring a high degree of resistance to HIV infection. Nevertheless, the mechanism of HIV replication is typically presented in textbooks at a level of abstraction that does not distinguish between the two cases (cf. Stine 2000, 64; Kalichman 1998, 16; Fan et al. 2000, 59). Let us turn now to a description of this mechanism.

HIV is an example of a retrovirus, which is so called because it reverses the normal flow of information from DNA to RNA. HIV replication proceeds according to the usual pattern for retroviruses. The genetic material of a retrovirus is encoded by RNA, and when a retrovirus infects a cell, its RNA serves as a template for the transcription of viral DNA, which is then insinuated into the cell's nuclear DNA. Once this occurs, the cell becomes a factory that produces HIV. The viral DNA integrated into the cell's genetic material codes for new viral RNA and proteins necessary for the functioning of the retrovirus. These materials are then assembled in the cytoplasm and new retroviruses bud from the cell membrane,



The mechanism by which HIV infects T-helper cells and macrophages is often depicted by diagrams like that in Figure 4.1 (cf. Kalichman 1998, 16; Stine 2000, 64). With the aid of such a diagram, it is possible to introduce more details about the mechanism than those just sketched above (cf. Kalichman 1998, 15–17; Stine 2000, Chapter 3). Glycoproteins (gp120) protruding from the surface of the HIV retrovirus attach to the T-helper cell at the CD4 (cluster determinant-4) receptor site. Note, therefore, that HIV infects only cells, such as macrophages and T-helpers, which display the CD4 receptor on their outer surface. Next, the viral RNA, ensconced in a protein coat, is injected into the host cytoplasm. Along with RNA, several enzymes necessary for the continuation of the infection are contained within the protein coat—most prominently, reverse transcriptase and integrase. The protein coat is quickly dissolved, and the viral DNA is then transcribed from the viral RNA by means of reverse transcriptase. The viral DNA is then integrated into the DNA of the host cell with the aid of integrase.

The cellular machinery of the host then proceeds to transcribe viral RNA and to synthesize proteins from the intruding DNA, thereby generating the materials needed to create new HIV viruses. These materials include new strands of viral RNA as well as several proteins and enzymes necessary for the functioning of the retrovirus. Besides reverse transcriptase and integrase, the enzyme protease is produced at this stage. Protease performs the role of splicing long protein strands into small, more usable pieces from which the internal protein coat can be constructed. Finally, the materials that constitute the new HIV virus assemble near the cell's external border, taking part of the host cell membrane with them as they bud forth. A large number of budding HIV viruses, therefore, kill the host

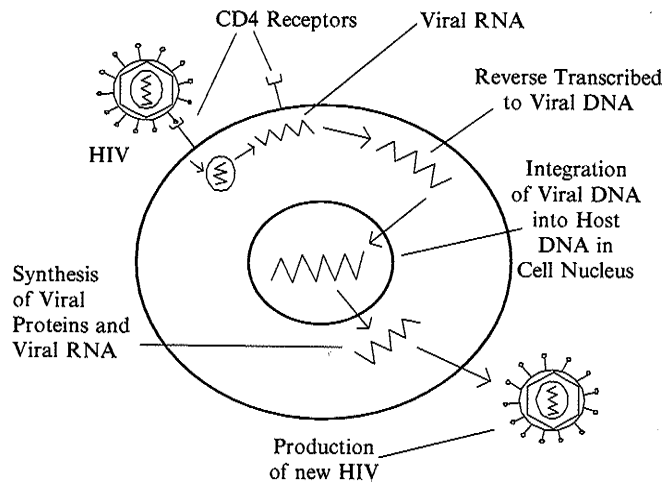



**Figure 4.1** HIV replication

cell by creating numerous perforations in its membrane.[3] Shortly after exiting the host cell, the virus constructs its inner structure of protein layers that enclose the RNA and vital enzymes. It is now ready to infect a new host.

Such is the mechanism by which HIV replicates. The evolutionary argument for why mechanisms can be identified with causal structure proposed in section 3.4 can be easily applied in this case. The HIV replication mechanism has clearly been honed by natural selection, which maintains the "normal" pattern described above as the statistically typical one. Moreover, the exasperating ability of HIV to evolve resistance to anti-retroviral therapies is a clear signal of the modularity of HIV replication. Thus, there is every reason to think that the HIV replication mechanism is capable of generating probability distribution and providing information concerning how those distributions will change, given interventions. In short, it is a causal structure.

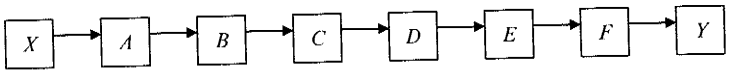
The HIV replication mechanism depicted in Figure 4.1 is also easily represented by a directed graph. For example, consider a collection of binary variables (1 = yes, 0 = no) defined as follows:

X: exposure to HIV
A: the virus attaches to the CD4 receptor
B: viral material enters the cytoplasm
C: reverse transcription of viral RNA occurs
D: viral DNA is integrated into host DNA
E: viral materials are produced by host cell
F: viral materials are assembled in preparation to form a new HIV virus
Y: a new infectious virus buds from the cell

Then we can represent the HIV replication mechanism with the graph in Figure 4.2. Of course, the mechanism could be represented in more or less detail, but this will suffice for present purposes.

## 4.2 FORMULATING THE PRINCIPLE

The disruption principle serves as a bridge from knowledge of mechanisms and things that interfere with them to qualitative conclusions about causal effects. Imagine that X is a positive causal factor for Y in the population P. Now suppose we ask whether there is a proper subset P′ of P in which the effect of X upon Y is *nullified*, that is, such that X is not causally relevant to Y within P′. The disruption principle provides a necessary and sufficient condition for the existence of such a subpopulation: for each member of P′; every mechanism from X to Y is blocked.





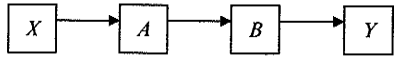




Figure 4.3 The flashlight mechanism

The disruption principle entails that if interventions on X alter the probability distribution of Y, there is at least one mechanism from X to Y. I will assume that directed graphs represent mechanisms as paths in which all of the arrows point in the same direction, as in the graph for the HIV replication mechanism given in Figure 4.2. Since that graph was rather lengthy, it will be convenient to use a shorter one for the purposes of illustration. For example, consider the simple case of pushing the "on" button of a flashlight. Pushing the button closes the electrical circuit, causing electricity to flow to the bulb, which in turn lights up. This causal chain could be represented by the graph in Figure 4.3, where X, A, B, and Y are binary variables representing, respectively, the button being pushed, the circuit being closed, electricity reaching the bulb, and the light shining. Thus the graph in Figure 4.3 depicts the mechanism of the flashlight.

A factor that nullifies the effect of X upon Y, then, must break this causal chain at some point. In the present example, such a factor is ready at hand; namely, the battery being dead. Let the variable Z represent the state of the battery: Z = 1 if the battery is charged and 0 otherwise. Adding Z to the graph from Figure 4.3, we have Figure 4.4. Notice that this graph does not indicate that X and Z interactively influence Y. That is, when Z = 0, X has no influence on Y; but this information is omitted by the graph in Figure 4.4. Indeed, the graph in Figure 4.4 could represent a situation in which A and Z influenced B independently of one another.[4]

Stating the disruption principle, then, is aided by a graphical notation for representing causal interactions in which an interfering factor severs a mechanism. I shall presume that each mechanism is represented by one directed path, such as that from X to Y in Figure 4.4. This is a purely terminological decision made for the pragmatic reason that it facilitates stating the disruption principle. Thus, several related causal chains that might be naturally referred to as a single mechanism would, in my terminology, be described as several interacting mechanisms. Factors that disrupt a mechanism, then, can be represented as follows. In the above example, there is a range of values of the variable Z (in this case

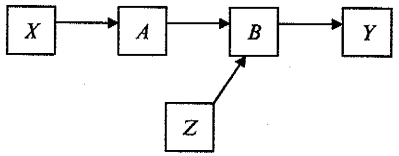



Figure 4.4 The flashlight mechanism and the battery

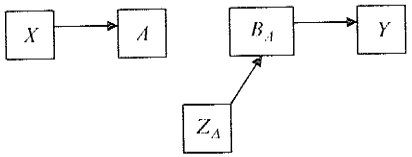



Figure 4.5 A disrupting factor

{0}) such that, when the value of Z is in that range, the effect of $A$ upon $B$ is nullified. I represent the elimination of the causal relationship between these variables by deleting the arrow joining them. Let $Z_\Delta$ indicate that the value of Z is such as to break the causal chain (the subscript delta standing for "disrupt"). In the example just described, $Z_\Delta$ would indicate that the battery is dead (i.e., $Z = 0$). I shall refer to variables with "$\Delta$" subscripts as *disrupting factors*. Then the causal structure for the subpopulation composed entirely of otherwise functional flashlights with dead batteries can be represented by the graph in Figure 4.5. The subscript $A$ appended to $B$ indicates that there is an arrow from $A$ to $B$ in the graph representing the causal relationships in the general population. Thus, given the graph in Figure 4.5, it is possible to unambiguously reconstruct the graph representing the causal relationships that hold in the general population, wherein the value of Z is not restricted to the disrupting set of values.

A disrupting factor, then, can be thought of as a switch that, when set to a particular position, breaks the mechanism connecting the cause and the effect. A more exact definition can be provided as follows. I shall use the expression *precedent variable* of a given node on the mechanism to refer to the directly prior node. For instance, in the mechanism represented in Figure 4.3, $A$ is the precedent variable on $B$. Suppose that $M$ is a mechanism through which $X$ influences $Y$. Let $V$ ($\neq X$) be a variable on $M$. Then

> *Definition 4.1*: Z is a *disrupting factor with respect to* $M$ just in case there is a variable $V$ in $M$ such that (1) Z is a cause of $V$, and (2) there is a range or interval $\Delta$ of values of Z such that when the value of Z is in $\Delta$, the variable in $M$ precedent to $V$ is not a direct cause of $V$.

For example, in Figure 4.4, Z is a cause of $B$, and when $Z = 0$, the variable on the mechanism precedent to $B$—namely, $A$—is not a cause of $B$. A disrupting factor Z will be said to be *active* with respect to a particular individual $p$ if the value of Z for $p$ is in the interval or range $\Delta$ that results in the mechanism being disrupted.

Since there may be several mechanisms connecting a given cause and effect in a population, it will be useful to speak of a *mechanism set* for a pair of variables. The mechanism set from X to Y in a population P consists of all of the mechanisms through which X influences Y that are found in at least one member of P. I shall use the notation $\mathbf{M}_{XY}$ to represent the mechanism set from X to Y. Mechanism sets are hence relative to a population. I shall use the expression "$\mathbf{M}_{XY}$ for $p$" to designate the subset



of $\mathbf{M}_{XY}$ that is instantiated in the individual $p$, where $p$ is any member of the population $P$ of concern. Then:

> *Definition 4.2*: $\mathbf{M}_{XY}$ for $p$ is *disrupted* if and only if, for each $M$ in $\mathbf{M}_{XY}$ for $p$, there is at least one disrupting factor that is active with respect to $p$.

Notice that, given definition 4.2, if $\mathbf{M}_{XY}$ for $p$ is empty, then it is trivial that $\mathbf{M}_{XY}$ for $p$ is disrupted. Let $\varphi_0$ be the relative frequency of individuals in P for which $\mathbf{M}_{XY}$ is disrupted. Then:

> *Disruption principle*: X is causally relevant to Y in P if and only if $\varphi_0 < 1$.

The disruption principle, therefore, links mechanisms to the probabilistic concept of causal relevance from Chapter 2. It may be helpful to quickly retrace these steps. According to the Machamer-Darden-Craver definition, "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). Chapter 3 argued that mechanisms so defined can, at least within the domain of molecular biology, be identified with causal structure. Causal structure, meanwhile, is defined as that which generates probability distributions and provides information concerning how those distributions change given interventions. Directed graphs, in turn, are one useful means of representing causal structure, and hence mechanisms if the two are identified. Finally, the disruption principle uses a slightly modified directed graph framework to state a relationship between mechanisms and the probabilistic concept of causal relevance defined in section 2.3.

Recall that X is causally relevant to Y just in case ideal interventions on X make a difference to the probability distribution of Y. As specified in definition 2.1, an ideal intervention is an exogenous cause that determines the value of the variable it targets. In the flashlight example, this could be something as simple as pushing the switch to "on." So, suppose our population P consists solely of flashlights with dead batteries. Then the graph in Figure 4.5 represents the disrupted mechanism found in each member of P, which means that $\varphi_0 = 1$. Therefore, by the disruption principle, X is not causally relevant to Y in P, that is, $P(Y = 1 \mid do(X = 1)) = P(Y = 1 \mid do(X = 0))$. In other words, when the battery is dead, moving the switch from "on" to "off" makes no difference to the probability that the light is shining. On the other hand, if there are some flashlights in P that possess the undisrupted mechanism, then $\varphi_0 < 1$, and hence the disruption principle entails that $P(Y = 1 \mid do(X = 1))$ is not equal to $P(Y = 1 \mid do(X = 0))$. Since $P(Y = 1 \mid do(X = 0)) = 0$, this entails that X is a positive causal factor for Y, that is, $P(Y = 1 \mid do(X = 1)) > P(Y = 1 \mid do(X = 0))$. The disruption principle, therefore, can link mechanisms to claims about positive causal relevance.

It should be noted that nullifying the effect of X upon Y does not necessarily entail determining the value of Y. For example, imagine

there is a gene that neutralizes the effect of smoking upon lung cancer. In the population of people possessing this gene, then, smoking would not be causally relevant to lung cancer. Nevertheless, members of this population might develop the disease through exposure to other carcinogens. On the other hand, neutralizing the effect of HIV in a population is sufficient for ensuring that nobody develops AIDS. Likewise, if the battery of the flashlight is dead, then the light does not shine. It is important to bear in mind, then, that this is a special feature of these two examples. It is not the case in general that eliminating the effect of $X$ upon $Y$ determines $Y$'s value.

Of course, the disruption principle is of little use unless some knowledge of relevant mechanisms and disrupting factors is available. Given the identification of mechanisms with causal structure, learning about mechanisms can be viewed as a special case of causal inference more generally conceived.[5] There are some discussions in the philosophical literature that examine strategies specifically suited for learning about mechanisms (cf. Bechtel and Richardson 1993; Darden 1991, 2002; Darden and Craver 2001, 2002). One such strategy, known as process tracing, is described in Chapter 5. Chapter 9 examines the relationship between process tracing and causal inference from statistical data. But for the moment, I set aside the concerns about how knowledge of mechanisms is to be acquired, assuming that the inquiry commences with some information on this score. Given such knowledge, the hunt for interfering factors can proceed by identifying points at which the mechanism is vulnerable to interference and searching for variables in the population capable of interfering with the mechanism at the specified points. Our knowledge of the mechanism need not be perfect for this hunt to commence, and as the example in the following section illustrates, the search for interfering factors can itself result in significant improvements in our knowledge of a mechanism.

### 4.3 RESISTANCE TO HIV INFECTION

Let us examine how the disruption principle comes into play in a realistic scientific example. Consider the question of whether there is a subpopulation in which the effect of exposure to HIV upon AIDS is nullified. It might seem that there is a straightforward solution to this problem that is independent of mechanisms: one need only find those who have been exposed to HIV but have not become infected. However, such a method, on its own, is an unreliable means for discovering subpopulations in which a causal effect is nullified, since there are several possible explanations for why the effect might not have followed the cause in a given case, including pure luck, exposure to a very mild form of the virus, or intrinsic resistance. The first two of these explanations yield very different predictions than the third about how the individual would



The disruption principle tells us that a fully resistant subpopulation, if it exists, is one in which every mechanism from HIV exposure to AIDS is severed in each individual. Hence, given the disruption principle, the search for the subpopulation in which the effect of HIV exposure upon AIDS is eradicated becomes the search for a disrupting factor, or set of disrupting factors, capable of blocking all mechanisms through which HIV brings about AIDS. Since all such mechanisms depend on HIV replication, that mechanism is a good place to look for such disrupters. That is, the set of mechanisms through which HIV produces the suite of symptoms associated with AIDS can be thought of as having the shape of a fan, with replication as its stem. Thus, since each mechanism shares this stem, blocking replication would sever all of them in one fell swoop. This thought is the point of the quotation at the head of this chapter. Let us turn, then, to the story of the discovery of a disrupting factor that seemed capable of nullifying the effect of HIV.

As described in section 4.1, HIV replication begins with the HIV retrovirus attaching to the CD4 receptor, which is exhibited on the surface of T-helper cells and cells of several other types, such as macrophages. However, the presence of the CD4 receptor is generally sufficient for an HIV virus to attach to a cell but not sufficient for the entry of the viral core into the cytoplasm (Maddon et al. 1986). Moreover, HIV strains that are capable of infecting macrophages are generally not able to infect noncirculating T-helper cells found in lymph nodes, and vice versa (cf. Gartner et al. 1986; Stine 2000, 141). Although these facts were recognized within a few years of the discovery of HIV,[6] an explanation of them was not forthcoming until nearly a decade later.

In 1996, it was discovered that distinct co-receptors present on macrophage and noncirculating T-helper cells play an important role in the entry of viral material into the host cell (Deng et al. 1996; Dragic et al. 1996). The co-receptor in the case of noncirculating T-helper cells is called CXCKR4 (X4 for brevity), and its counterpart for macrophages is known as CC-CKR5 (R5 for brevity).[7] M-tropic HIV utilizes R5, while T-tropic HIV depends upon X4, thereby accounting for the difference in affinities of the two strains.[8] Within the same year, it was discovered that some individuals who had not become infected with HIV despite repeated exposures possessed a mutation that inhibited the normal R5 co-receptor (Samson et al. 1996; Liu et al. 1996). When exposed in vitro to M-tropic HIV strains, cells from these individuals

> ...required about 1000-fold more virus to establish infection than control cells from unexposed donors. While a small fraction of the cells did become infected with this high inoculum, the virus failed to replicate further. (Liu et al. 1996, 367)

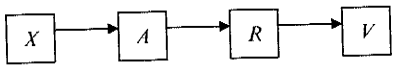
As noted in the foregoing section, M-tropic HIV predominates in the early and asymptomatic stages of infection. Thus, if replication of M-tropic HIV is blocked, the progression of the infection is strongly, if not completely,

inhibited. The absence of the normal R5 co-receptor was linked to a homozygous mutation, in which thirty-two base pairs in the ordinary gene coding for the co-receptor were deleted. As this mutation appears to produce no other abnormal phenotypic effect, it is a veritable genetic blessing for those lucky enough to have inherited it. The homozygous mutation was estimated to occur among approximately 1 percent of "persons with western European heritage" (Liu et al. 1996, 373; Dean et al. 1996). The heterozygous condition, which appears to confer a more attenuated resistance (Eugen-Olsen et al. 1996), is surprisingly common—with estimates ranging from about 20 percent (Liu et al. 1996, 373) to 14 percent (Dean et al. 1996, 1860) among Caucasians. The mutant allele was not found in African or Asian populations (Samson et al. 1996, 722). There was a striking negative association between HIV infection and the homozygous mutation. In several large data sets, *all* of those homozygous for the thirty-two-base-pair deletion were HIV negative (Samson et al. 1996, 722; Dean et al. 1996, 1860). These data stimulated hope that the homozygous mutation affecting the R5 co-receptor might confer complete resistance to AIDS.[9]

The thread of this scientific detective story will be taken up again in Chapter 7, so for now let us consider what, if the disruption principle is true, would have to be the case for the hope just described to be realized. Consider the segment of the M-tropic HIV replication mechanism that is disrupted by the homozygous mutation affecting the R5 co-receptor, which is represented in Figure 4.6.

As in Figure 4.2, $X$ and $A$ are binary variables indicating exposure to HIV and attachment of HIV to the CD4 receptor, respectively, while $R$ is a binary variable indicating attachment to the R5 co-receptor and $V$ represents the rate of reproduction of M-tropic HIV. Suppose that the presence of the homozygous thirty-two-base-pair deletion fully blocks attachment to the R5 co-receptor. Then we have Figure 4.7. Here $Z$ is a variable representing the presence of the mutation affecting the R5 co-receptor that takes three values {homozygous normal; heterozygous; homozygous mutant}. The subscript "$\Delta$" in this case indicates that $Z$ takes on the third of these values. Thus, given the supposition that the homozygous mutation completely blocks the mechanism in Figure 4.6, it also fully blocks M-tropic HIV reproduction if there is no path from $X$ to $V$ that circumvents $R$.

But even if this is so, it would not necessarily follow that the homozygous mutation confers immunity to HIV infection and AIDS, since T-tropic HIV does not utilize the R5 co-receptor. However, given that M-tropic HIV predominates in the early stages of HIV infection, it is possible that the continuation of infection by T-tropic HIV depends





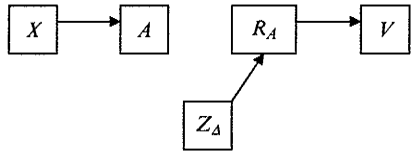




**Figure 4.7** The homozygous mutation

upon the replication of M-tropic strains. Letting $V_\Delta$ indicate $V = 0$, this thought can be represented by the graph in Figure 4.8. In this graph, $S$ is a binary variable indicating entry to the cytoplasm for T-tropic HIV, and $T$ is some unspecified stage of T-tropic HIV replication that is blocked by the failure of the M-tropic HIV infection. Precisely what $T$ might consist of depends on how the absence of M-tropic replication inhibits that of T-tropic, an issue that will be taken up in section 7.1.

From the graph in Figure 4.8, it can easily be seen that, given the disruption principle, the homozygous mutation inhibiting the R5 co-receptor completely nullifies the effect of HIV exposure upon AIDS if and only if there is no path from HIV exposure to AIDS that bypasses both $R$ and $T$. In other words, it must be that the homozygous mutation completely blocks the M-tropic HIV replication mechanism (or the set of them, if there are several), *and* there is no mechanism from HIV exposure to AIDS that bypasses replication of M-tropic HIV. If it were to be discovered that the mutation inhibiting the R5 co-receptor did *not* confer complete immunity, the disruption principle would entail that at least one of these two conditions is false.

This example illustrates how the disruption principle captures a relatively commonsense inference concerning mechanisms and nullified causal effects. But if the role of the disruption principle were limited to reconstructing such examples, there would hardly seem to be much point in taking the time to provide a precise articulation of it. However, there is a twofold value in clearly stating and highlighting the disruption principle. First, as will be seen in Chapter 6, there are less obvious consequences of the disruption principle regarding extrapolation of causal

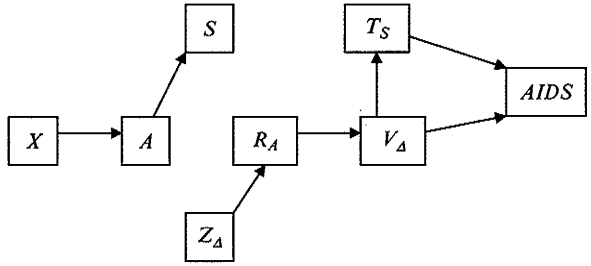



**Figure 4.8** How the homozygous mutation might confer HIV immunity

claims. Second, a clear statement of the principle facilitates a careful examination of the circumstances in which it is and is not a reasonable assumption, a topic addressed in the remainder of this chapter.

## 4.4 WHY BELIEVE THE DISRUPTION PRINCIPLE?

I hope that the plausibility of the disruption principle has been motivated by the HIV example examined above. Nevertheless, it is worth considering whether the principle is supported by other more general, and more familiar, principles connecting causal structure and probability. In fact, granting that mechanisms are identified with causal structure, the disruption principle is a straightforward consequence of two propositions known as the *Principle of the Common Cause* (PCC) and the *Faithfulness Condition* (FC). This result is of interest in two respects. First, it implies that any justification for the PCC and the FC is also a justification for the disruption principle. Second, it suggests that circumstances in which the PCC or FC fails may be ones in which the disruption principle fails as well. Since alleged counterexamples have been raised against both the PCC and the FC, this last observation is far from being an idle point. However, I show that the PCC is on very firm ground in the type of experimental context that is of concern here. The case of the FC is more complex. I argue that the FC is reasonable for heterogeneous populations—such as naturally occurring biological populations—but not necessarily for extremely homogeneous ones, such as closely inbred strains of laboratory mice.

### 4.4.1 The Disruption Principle and the PCC

Let us begin by considering the connection between the disruption principle and the PCC.[10] The PCC can be stated in the following way.

> *PCC:* For any two distinct variables $X$ and $Y$, if $X$ and $Y$ are not causally connected, then they are probabilistically independent.

Two variables are *causally connected* just in case one is a cause of the other or there is a common cause of both. Thus, the PCC says that two variables are probabilistically dependent only if one is a cause of the other or there is a third variable that is a common cause of both.

Given the identification of mechanisms with causal structure defended in Chapter 3, one half of the disruption principle is a direct consequence of the PCC. That is, the disruption principle is a biconditional that, in one direction, asserts: If there is no undisrupted mechanism from $X$ to $Y$ in the population P, then $X$ is not causally relevant to $Y$ in P. By definition 2.3, $X$ is causally relevant to $Y$ just in case $X$ and $Y$ are probabilistically dependent under ideal interventions on $X$. But in the context of an ideal intervention on $X$, $Y$ does not cause $X$ and there is no common cause of $X$ and $Y$. Thus, if there is no undisrupted mechanism from $X$ to $Y$, then $X$



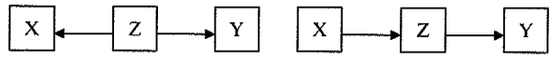



**Figure 4.9** An illustration of the causal Markov condition

PCC, therefore, it follows that $X$ and $Y$ are probabilistically independent in such circumstances.

The PCC is itself a consequence of a more general principle connecting causality and probability known as the *causal Markov condition* (CMC). Roughly, the CMC asserts that, conditional on its direct causes, any variable is probabilistically independent of any set of other variables that do not include its effects. The CMC, therefore, entails the familiar "screening-off" rule. For example, consider the two directed graphs in Figure 4.9. If these graphs satisfy the CMC, then $X$ and $Y$ are probabilistically independent, conditional on $Z$ in both.

Probably the most common basis provided for the CMC is that it is true of acyclic, deterministic causal structures in which the exogenous variables are probabilistically independent (cf. Pearl 2000, 30; Spirtes, Glymour, and Scheines 2000, 32; Glymour 2001, 27).[11] This proposition can be extended to indeterministic causal structures (Steel 2005), leaving only the other two assumptions—probabilistic independence of exogenous variables and absence of causal cycles—as matters of concern. The disruption principle asserts, in part, that if there are no undisrupted mechanisms from $X$ to $Y$, then ideal interventions on $X$ make no difference to the probability of $Y$. Recall that an ideal intervention is exogenous, that is, it is neither an effect of, nor shares a common cause with, any of the variables being studied (in this case, $X$ and $Y$). The best way to ensure that this condition is satisfied in practice is to assign the value of the targeted variable ($X$, in this case) on the basis of some random process, such as tossing a coin.

So, consider the relationship between $X$ and $Y$, when the values of $X$ are randomly assigned by an ideal intervention, which is represented by the variable $I$. It is easy to show that, in the causal structure relating only $I$, $X$, and $Y$, all exogenous variables are probabilistically independent of one another *and* there are no causal cycles. From items (a) and (b) of definition 2.1, we know that $I$ is the sole cause of $X$ and a direct cause only of $X$. Moreover, item (c) of definition 2.1 asserts that $I$ is exogenous. Consequently, the only two possible causal structures relating $I$, $X$, and $Y$ are those represented in Figure 4.10. Since there are no causal cycles in either case, the requirement that the causal structure be acyclic is satisfied. It is trivial that every exogenous variable is probabilistically independent of every other in the graph on the left, since in that graph there is only one exogenous variable, $I$. In the graph on the right, there are two exogenous variables, $I$ and $Y$. But given randomization, we know that the intervention $I$ is probabilistically independent of every other exogenous variable. Hence, $I$ and $Y$ are probabilistically independent in that graph. Thus, in

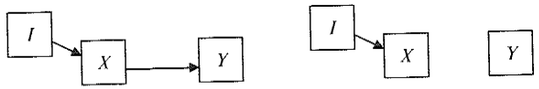



Figure 4.10 Two ideal interventions

both graphs the exogenous variables are probabilistically independent. Consequently, it follows that the CMC, and hence the PCC, is true in any case involving two variables, one of whose values is randomly assigned by an ideal intervention. Since randomization is the standard experimental procedure for ensuring that an intervention is exogenous, the half of the disruption principle asserting that a causal effect is nullified when all mechanisms are blocked is on firm ground in the context of experiments.[12]

Of course, this does not show that there are no practical challenges confronting applications of the PCC in the present context. For example, the statistical problem of reliably drawing inferences concerning probabilities on the basis of data in a sample is ubiquitous. The presence or absence of a statistically significant correlation coefficient in the data may be the result of mere chance. In addition, it may be difficult to know whether an actual experiment satisfied the conditions of an ideal intervention. But although they are real, these challenges are independent of the PCC; they are general problems for statistical inference and experiment.[13]

### 4.4.2 Genetic Redundancy and the Faithfulness Condition

Consider the relationship between exercise and weight loss. Additional exercise results in more calories being burned, but it also stimulates one's appetite. Letting the variables *E*, *A*, and *W* denote exercise, appetite, and weight, respectively, this situation can be represented by the graph in Figure 4.11. Conceivably, the strengths of these two paths from exercise to weight could exactly cancel out and make *E* and *W* probabilistically independent, thereby contradicting the FC. Nevertheless, it is clear that modern medicine does not take this possibility seriously: physicians and others endeavoring to promote public health have long encouraged overweight people to get more exercise.

Peter Spirtes, Clark Glymour, and Richard Scheines (hereafter, SGS) prove that though the exact balancing of strengths of counteracting causal

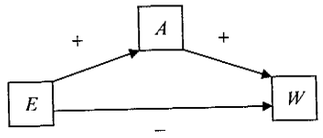



Figure 4.11 Counteracting causal paths from exercise to weight



paths is conceivable, given certain apparently plausible assumptions, it is monstrously improbable. To see the idea, consider the following linear causal model. In this model, *a*, *b*, and *c* are linear coefficients representing quantitative strength of influence. The subscript, lowercase *e*'s are called "error terms," and represent any source of variation in the dependent variable not accounted for by its direct causes. It is assumed that error terms are independent and are normally distributed with zero means. A *parameterization* of a linear causal model consists in specifying numerical values for the coefficients and for the variances of the error terms. Notice that *X* and *Z* would be uncorrelated, and the FC would be false, for any parameterization in which $b + ac = 0$.

$$X = e_x$$
$$Y = aX + e_y$$
$$Z = bX + cY + e_z$$

SGS's theorem states conditions under which parameterizations that result in such precise canceling out have probability zero.[14] Take any model in which effects are linear functions of their causes. Suppose that this model contains $n$ parameters. For example, $n = 6$ in the linear causal model. Consider the $n$-dimensional space of all parameterizations of this model, that is, each point in the space corresponds to a parameterization. Now consider any subset of that space consisting solely of parameterizations that violate the FC. In the model, an example of such a subset would be one in which every parameterization makes $b + ac = 0$. Then it can be shown that any subset of the $n$-dimensional space containing *only* parameterizations that violate the FC is of $n - 1$ dimensionality or less. Then the following assumption is made:

> **L:** In an $n$-dimensional space of parameterizations, any subset of $n-1$ dimensionality or less has probability zero.[15]

Thus, it follows that any subset of the space of parameterizations of a linear causal model containing only parameterizations that violate the FC has zero probability.

However, not everyone regards SGS's theorem as a compelling motivation for the FC, and some have argued that exceptions to the FC are not uncommon. For example, according to Cartwright:

> Faithfulness will be violated if the two processes are equally effective and cancel each other out. It is not uncommon for advocates of DAG-techniques to argue that cases of cancellation will be extremely rare, rare enough to count as non-existent. That seems to me unlikely, both in the engineered devices that are sometimes used to illustrate the techniques and in the economic and medical cases to which we hope to apply the techniques. For these are cases where means are adjusted to ends and where unwanted side effects tend to be eliminated wherever possible, either by following an explicit plan or by less systematic fiddling. (1999, 118)

A similar argument is made by Kevin Hoover.

> Spirtes et al. (1993, 95) acknowledge the possibility that particular parameter values might result in violations of faithfulness, but they dismiss their importance as having "measure zero." But this will not do for macroeconomics. It fails to account for the fact that in macroeconomic and other control contexts, the policymaker aims to set parameter values in just such a way as to make this supposedly measure-zero situation occur. To the degree that policy is successful, such situations are common, not infinitely rare. (2001, 171)

Cartwright and Hoover both make it clear that they do not mean to say that the FC is *always* false, but only that it fails in certain situations, namely, those in which there is some process that selects for canceling out causal paths. For instance, in the exercise-weight example it seems unlikely that there is selection in favor of precisely counterbalancing parameterizations. Hence, their argument would not support the conclusion that exceptions to the FC are probable in that case. However, they do think that there is often selection for counterbalancing paths, which would imply that the FC is problematic as a general principle.[16]

If this objection is right, then there must be an assumption of SGS's theorem that is false in circumstances of the sort Cartwright and Hoover indicate. It is easy to see that the assumption called into question by Cartwright and Hoover's line of argument is **L**.[17] They claim that when there is selection for parameterizations in which causal paths cancel out, it is, for instance, probable in the linear model above that $b + ac = 0$. But the subset consisting solely of parameterizations that make $b + ac = 0$ is a two-dimensional subset of the three-dimensional parameter space. Hence, if it is probable that the actual parameterization is within that subset, then it is false that the probability of every subset of $n - 1$ dimensionality or less is zero.

Indeed, it would be unreasonable to maintain that $n - 1$ dimensional subsets of $n$-dimensional spaces must *always* have zero probability. For example, such a claim would entail that we must be certain a priori that no quantity is equal to any other quantity. This point can be appreciated by reference to the diagram in Figure 4.12. In the diagram, the subset of pairs of values in which $a$ equals $b$ is represented by the diagonal line in the square, which of course is one dimension less than the two-dimensional plane. Consequently, SGS's theorem can serve as a motivation for the FC only provided there is some explication of the conditions under which **L** is true and of why we should think that those conditions hold in the domain of application of the FC.[18] I suggest that **L** is a reasonable assumption when parameter values are affected by a large number of uncontrollable factors, a situation which is the norm in heterogeneous populations.

Consider a social planner attempting to do what Hoover describes,



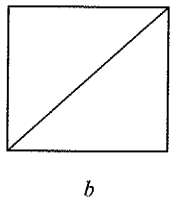



**Figure 4.12** The Subset in which $a = b$

undesired side effect of some policy. For example, imagine a road improvement program that involves resurfacing and widening a number of large thoroughfares and some smaller side streets. Although improved, safer roads contribute to fewer traffic accidents, they also have the unfortunate side effect of increasing speeding, which is a significant cause of traffic fatalities. Letting $R$, $S$, and $T$ be variables denoting road improvement, rates of speeding, and traffic fatalities, respectively, the example is represented by the graph in Figure 4.13. Suppose that, initially, the net effect of the road improvement is to increase the rate of traffic fatalities. To offset this problem, more police are hired to patrol the newly improved roads and the fines for speeding are increased. However, given a tight budgetary situation, the social planners do not want to spend more money on speeding prevention than necessary. They want to do just enough to make the two causal paths cancel out, and no more.

The strategy of the social planners in this case is to implement changes in the situation that will weaken the positive influence of $R$ upon $S$ so as to even the balance between the two paths. In principle, if the strength of influence of $R$ upon $S$ can be fine-tuned independently of the other parameters, this would be possible. But the relevant question is whether the social planners really can reliably make the exact canceling out occur, or at least be sufficiently approximated for practical purposes. Their ability to do so depends on being able to establish the following two things:

> *Selection of Parameters*: A process that tends to concentrate the weight of the distribution of parameterizations on a subset in which the FC is violated.

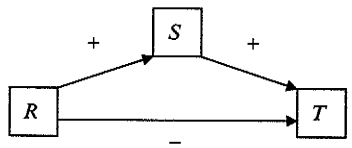



**Figure 4.13** Road improvement and traffic fatalities

*Homogeneity of Parameters*: The absence of factors that perturb parameter values and thereby alter their distribution in uncontrolled ways.

If selection and homogeneity were perfectly accomplished, then the entire distribution of parameterizations would be restricted to an $n - 1$ dimensional subset of the parameter space. With regard to the graph in the linear causal model, an example of this would occur when it is certain that the parameterization makes $b + ac = 0$. Thus, when selection and homogeneity are perfectly satisfied, assumption **L** of SGS's theorem (that all $n - 1$ dimensional subsets of the parameter space receive probability zero) is false. Of course, it is unlikely that selection and homogeneity will be perfectly fulfilled in real-life examples. But if the distribution of parameterizations were tightly focused on an $n - 1$ dimensional subset of the parameter space rather than completely restricted to that subset, very near violations of the FC would be probable even if **L** were true. Moreover, there is little practical difference between *strict* and *very near* failures of the FC, since in either case, causal connections fail to give rise to correlations that are detectable in any obtainable sample size. Thus, Cartwright and Hoover's objection would be vindicated if very near exceptions to the FC were common in the principle's intended domain of application.

The above considerations show that selection and homogeneity suffice to make (near) exceptions to the FC probable. And Cartwright and Hoover's objection points out that it is not rare that someone or something endeavors to put a selection process in place. Trying and succeeding, however, are two very different things. In the road improvement example, it was assumed that the policymakers endeavored to make the causal paths cancel out through an adjustment of the strength of influence of $R$ upon $S$. Bringing about the desired balance, then, requires knowing the requisite value of this parameter and being able to fine-tune it accordingly. Yet it is far from clear that this knowledge or ability is typically possessed by policymakers. How many additional police cruisers patrolling the streets would be required to reduce the value of the parameter by a given amount, for instance? For the moment, let us put aside this concern and suppose that the policymakers can devise a selection process.

Even if there is a process at work that tends to focus the probability distribution of parameterizations around an FC-violating subset, it does not follow that exceptions or near exceptions to the FC are probable, since the distribution of parameters might also be influenced by other trends that undo the work of the selection process. Suppose that there is a wide variety of difficult-to-predict or -control factors at play that are capable of altering the values of the parameters (i.e., that homogeneity does not hold even approximately). Clearly, these disturbing factors would be expected to increase the variance of the distribution of parameterizations, thus increasing the chance that the actual parameterization would fall in a



region distant from a subset in which the FC is false. In addition to enlarging the variance of the distribution, factors that alter the values of parameters can also change its mean if not all parameters are uniformly susceptible to disturbance. For instance, if some parameters are more susceptible than others to factors that alter their values in a particular direction, then the mean of the distribution might be driven away from an FC-violating subset. In the road improvement example, if the effect of $R$ upon $S$ is sensitive to factors that tend to increase its value while the other parameters are relatively stable, then the mean of the parameterizations will move toward a positive net effect of $R$ upon $T$.

The simple moral, then, is that the existence of a selection process can fail to make exceptions or near exceptions to the FC probable when a variety of uncontrollable factors that perturb parameter values is present.[19] Consequently, noting the presence of a selection process does not suffice to show that (near) violations of the FC are likely to occur. Yet Cartwright and Hoover's objection merely points out that it is common for selection processes to be present or at least for some effort to be made to create them, and then concludes that exceptions or near exceptions to the FC are likewise commonplace. Consequently, their argument is invalid on two grounds. First, effectively designing and implementing a selection process may be very difficult, so the fact that there is some effort afoot to create a selection process provides little assurance that one exists.[a] Second, even if a selection processes were common, Cartwright and Hoover's conclusion would follow only when homogeneity obtains. Yet it is obvious that the opposite is typically the case for the heterogeneous populations that are the concern of this book. Causal relationships in biological and social phenomena generally depend upon variable factors that are difficult to predict or control. Hence, the intended domain of the FC for the present purposes consists of causal systems of which it is quite doubtful that homogeneity is typically true or even approximately true.[20]

In short, Cartwright and Hoover's objection has failed to show that exceptions or near exceptions to the FC are common in its intended domain of use. Nevertheless, it would be a mistake to conclude that the FC is *always* an unproblematic assumption with regard to complex systems. In particular, near violations of the FC are probable when both selection and homogeneity are approximated, and it is arguably the case that this situation is not infrequent in gene knockout experiments.

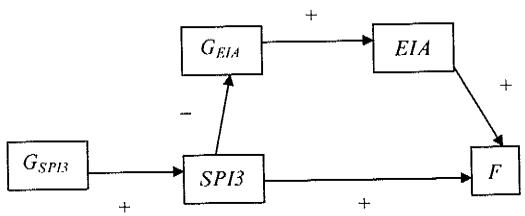
Although it is not an example that they discuss, genetic redundancies illustrate the type of situation in which Cartwright and Hoover claim that exceptions to the FC are probable. For example, imagine a gene that serves as a template for the transcription of a protein that normally performs a specific set of functions in a cell, but when that protein is not present in sufficient quantities, the transcription of a distinct yet functionally similar protein from a second gene is increased. Moreover, it is plausible that there would be an adaptive benefit in the quantitative strengths of the two paths exactly counterbalancing one another. For example, maintaining

the function may require that the sum of two products be kept within fairly narrow bounds. Hence, it would not be optimal for both genes to normally be transcribed together, while it would be beneficial that the function be maintained at the normal rate when the usual product is not present in adequate quantities. Thus, natural selection would constitute a process that favors parameterizations in which the counteracting paths exactly or very nearly cancel out.[21]

Moreover, apparent near exceptions to the FC are not rare in gene knockout experiments. For instance, a recent gene knockout study (Scarff et al. 2004) examined a particular protease inhibitor, SPI3, believed to have several important functions which primarily involve preventing certain proteases from affecting nontarget cell and tissue types. The investigators produced a strain of mice in which the gene from which SPI3 is transcribed was disabled, but surprisingly this mutant strain appeared completely normal and showed no apparent difference in any of the several functions to which SPI3 is believed to be relevant. Given the FC, this result would constitute strong evidence that SPI3 is not a cause of any of the functions in question. However, that was not the conclusion drawn by the researchers. They noted that among mice in which the SPI3 gene had been knocked out, the presence of a second protease inhibitor, EIA, was increased. Since EIA is functionally similar to SPI3, this suggested that the failure of the gene knockout to produce any detectable difference between the mutant and wild-type strains could be explained by a compensating pathway.

The authors suggested two possible mechanisms through which the knockout of the gene for SPI3 could stimulate the increased transcription of EIA (ibid., 4080). In both cases, higher levels of SPI3 inhibit the transcription of the gene from which EIA is synthesized, thereby suppressing EIA under normal circumstances. The basics of the hypothesis, then, can be represented in the graph in Figure 4.14. The variables $G_{SPI3}$ and $G_{EIA}$ represent the rate of transcription of the genes for SPI3 and EIA, respectively. According to the authors, the results of their experiment "indicate that EIA levels are increased in SPI3-deficient mice to compensate for the loss of SPI3" (ibid., 4079).

Nor is the above example an aberration.[22] As Sandra Mitchell (2003, 154–55) notes, redundancy is a common challenge for gene knockout






experiments. Indeed, a recent issue of *Nature* included a news feature on exactly this topic. It pointed out that:

> In many cases, a mutant mouse [produced by the gene knockout] does not show any obvious characteristics—or phenotype. In others, the phenotype disappears when the disabled gene is crossed into a different strain of mouse. Indeed, clear and consistent phenotypes now seem to be the exception rather than the rule. (Pearson 2002, 8)

A common explanation given for this state of affairs is genetic redundancy: "These results often reflect the fact that genes acting in parallel pathways can compensate for the one that is missing" (ibid.).[23] Moreover, the journal *Molecular and Cellular Biology* has, since November 1999, dedicated a section of each issue to the topic of gene knockout studies that find surprising differences, or lack of differences, between mutants and controls.[24] That was where the study described above was published.

As explained above, the presence of a selection process alone is not sufficient to make exceptions or near exceptions to the FC probable; homogeneity is also required. But this latter condition is much more likely to be approximated in the context of a gene knockout experiment than in, say, a wild population of mice. The mice in knockout experiments are typically generated from extremely genetically homogeneous strains that have been reared for numerous generations under standard laboratory conditions. The procedure by which knockout strains are generated further enforces this homogeneity. Knockout strains of mice are generated by disabling the target gene in embryonic stem (ES) cells and then either injecting these ES cells into a blastocyst or aggregating them with an embryo at an earlier stage.[25] Since they are the product of cells from more than one individual, such modified embryos are known as chimeras. Chimeras formed by blastocyst injection or aggregation at earlier stages will, if viable, transmit the modified genes through the germ line. The knockout strain can then be generated from the mutant germ line cells selected from the chimeras (if these can be successfully created). The knockout strains, therefore, are generated not only from strains that are quite genetically homogeneous, but also in a way that produces an effective genetic bottleneck, since the knockout strain ultimately derives from the chimeras, which descend from the modified ES cells and their zygote hosts.

Gene knockout experiments, then, are a context in which it is not unlikely that both selection and homogeneity are satisfied to a reasonably good approximation, and hence near exceptions to the FC are probable. This analysis of failures to find effects in gene knockout experiments has an empirical consequence. In cases in which there is good reason to believe that a causal mechanism is present, despite the null result of a knockout experiment, it is likely that a probabilistic dependence between the suspected cause and effect will appear if the conditions of the experiment are varied. And in fact, surprising absences of difference between

mutants and controls in one gene knockout experiment often emerge in other mouse strains, or strains reared in distinct environmental conditions (Pearson 2002, 8–9).

That the FC is a problematic assumption in the case of gene knockout experiments does not show that it is an inappropriate assumption generally. For example, the challenge to the FC arising in gene knockout experiments should not be expected to transfer to studies involving more heterogeneous populations, such as human subjects, for the reasons discussed above. Thus, the lack of genetic and environmental homogeneity among experimental subjects, surprisingly enough, can facilitate the discovery of causal structure in virtue of providing a more hospitable setting for the FC. Of course, this is not to deny that there are genuine benefits of uniform populations of experimental subjects, but it does show that experiments involving such subjects also have some potential downsides. It is not too difficult to see how those downsides can be avoided in the present case: vary the genetic and environmental backgrounds of the experimental populations. But although this prescription is simple enough in principle, there are practical obstacles to implementing it in the laboratory. As the news feature from *Nature* cited above observes:

> Ideally, experiments on knockout mice would routinely include work on multiple strains. In practice, most researchers in the field ague that this is not realistic—creating a single knockout strain can take up the majority of a three-year PhD project. (Pearson 2002, 8)

Barring the development of methods that allow knockout strains to be created more easily, the FC seems likely to remain a problematic assumption in gene knockout experiments for the foreseeable future.

So, where does all this leave the disruption principle? Recall that the disruption principle has two parts. First, if there is no undisrupted mechanism from $X$ to $Y$, then ideal interventions on $X$ do not alter the probability distribution of $Y$. As explained in section 4.4.1, this part rests on solid ground. Second, the disruption principle asserts that if there is an undisrupted mechanism from $X$ to $Y$, then ideal interventions on $X$ alter the probability distribution of $Y$. In this section I have argued that, some objections notwithstanding, this is normally an appropriate assumption with respect to complex causal systems wherein the strengths of causal influences are subject to a wide array of uncontrolled factors. Nevertheless, there are some situations, as illustrated by gene knockout experiments, in which this heterogeneity is substantially reduced, and hence in which near exceptions to the disruption principle are more probable.

## 4.5 CONCLUSION

This chapter introduced, illustrated, and explored the range of applicability of the disruption principle, a central premise of the mechanisms approach to extrapolation. It was shown how this principle could be



stated by way of the formalism of directed graphs, and its role was illustrated by reference to an example drawn from HIV research. It was also shown that the principle is a logical consequence of two other, more familiar propositions connecting causality and probability: the Principle of the Common Cause (PCC) and the Faithfulness Condition (FC). The aspect of the PCC specifically relevant to the disruption principle was shown to be unproblematic, but the case of the FC was more complex. Although there is a good motivation for the FC for heterogeneous, naturally occurring biological populations, that justification does not extend to extremely homogeneous populations, such as the strains of laboratory mice typically used in gene knockout experiments. This result delineates more exactly the conditions under which the FC is and is not an appropriate methodological principle of causal inference, and it shows that the experimental practice of holding all background factors fixed is not always a virtue in the study of complex systems, since such experimental arrangements may conceal probabilistic dependencies between cause and effect that are present in messier, real-world populations.

# 5

# Extrapolation, Capacities, and Mechanisms

Imagine that a chemical occurring in some consumer products has been found to be carcinogenic if administered in large doses in rats, and the question is whether it is also a carcinogen in humans. The mere repetition of the experimental result in rats is not sufficient to answer this question, since the physiology of rats may differ in some relevant respect from that of humans. This is an example of extrapolation: given some knowledge of the causal relationship between $X$ and $Y$ in a base population, we want to infer something about the causal effect of $X$ upon $Y$ in a target population.[1] For instance, in the example above, we know that the chemical is a positive causal factor for cancer in rats and we want to know whether it is also such in humans. Difficult cases of extrapolation are ones in which the base and target populations may differ in relevant respects and, moreover, in which ethical or practical considerations prohibit directly testing the claim at issue by experiment in the human target population.

The most straightforward way to approach extrapolation is to presume that what is true of one population is also approximately true of other related populations unless there is some specific reason to think otherwise. I call this inferential strategy *simple induction*. However, since simple induction would inevitably lead to many mistaken extrapolations, a more sophisticated approach would be highly desirable. Any account of extrapolation that goes beyond simple induction must confront two basic challenges. The first is what I call the *extrapolator's circle*. Simple induction relies on some criterion of relatedness, such as phylogeny or type of economic system. The shortcomings of simple induction stem from the fact that satisfying such criteria is often not sufficient for being a reliable basis for extrapolation. Consequently, additional information about the similarity between the model and the target—for instance, that the relevant mechanisms are the same in both—is needed to justify the extrapolation. The extrapolator's circle is the challenge of explaining how we could acquire this additional information, given the limitations on what we can know about the target. In other words, it needs to be explained how we could know that the model and the target are similar in causally relevant respects without already knowing the causal relationship in the target. The second challenge arises from the inevitable presence, in the biological and social sciences, of causally relevant differences between the model and the target. Thus, any adequate account of extrapolation in



heterogeneous populations must explain how extrapolation can be possible even when such differences are present.

I argue that existing accounts of extrapolation fail to answer these challenges. One proposal I consider maintains that capacities or causal powers that exert a characteristic influence independently of context are a basis for extrapolation. However, this proposal does not adequately explain how one is to know that one is dealing with a capacity rather than a context-sensitive causal relationship, aside from already having found that the causal relationship obtains in all of the contexts in question. Thus, without some further elaboration, the capacities proposal does not address the two challenges just described. A mechanisms approach to extrapolation could be regarded as such an elaboration of the capacities proposal or as a separate approach. According to this approach, knowledge of mechanisms linking cause and effect and knowledge of factors capable of interfering with these mechanisms can significantly facilitate extrapolation. As noted in Chapter 1, this idea has been suggested by several philosophers, social scientists, and biologists. However, a mere invocation of mechanisms does not resolve the extrapolator's circle, nor does it explain how extrapolation can be possible in the face of causally relevant disanalogies. The mechanism approach needs to explain how the suitability of the model could be established without already knowing all of the important details about the mechanism in the target. Moreover, some differences in the mechanism in the model and the target are inevitable in biology and social science. Thus, the mechanisms approach requires an account of how extrapolation can be justified even when such differences are present.

In this chapter, I develop a more satisfactory version of the mechanisms approach to extrapolation. The central concept is a mode of inference I call *comparative process tracing*, which aims to assess the suitability of the model as a basis for extrapolation. Comparative process tracing depends upon background information concerning likely similarities and differences between the model and the target. If significant differences between the model and the target are likely to be restricted to a relatively small number of stages of the mechanism, then comparisons at those stages may provide good grounds for the suitability of the model. The number of stages that must be compared can be reduced further if upstream differences must result in differences at an observable downstream point in the mechanism. Moreover, knowledge of just a few stages of the mechanism in the target a*lone* (that is, without knowledge of the model) might not suffice for firm conclusions regarding the existence of a mechanism in the target. Hence, my proposal provides an analysis of how extrapolation can be justified despite the extrapolator's circle while indicating conditions in which it is a genuine problem. I illustrate my account of comparative process tracing with a case study concerning the carcinogenic effects of aflatoxin $B_1$. The question of how useful the mechanisms approach to extrapolation developed here is likely to be in social science is taken up in Chapter 8.

The aflatoxin example also illustrates how extrapolation may be justifiable even when there are some causally relevant disanalogies between the model and the target. I argue that the closeness of the match between model and target required for extrapolation depends upon the specificity of the causal claim to be extrapolated. While similarity in all causally relevant respects may be required for extrapolating an exact, quantitative causal effect, it is not required for extrapolating qualitative causal claims. In particular, claims about positive or negative causal relevance can be extrapolated even when there are causally relevant disanalogies. This point is illustrated by the aflatoxin example, wherein a causally relevant difference between the animal model and the human suggests that the carcinogenic effect is less in the model than in humans. Yet this difference does not indicate that it would be a mistake to extrapolate the claim that exposure to aflatoxin $B_1$ increases the chance of liver cancer in humans. A more general and precise characterization of conditions that suffice for extrapolating claims about positive and negative causal relevance is given in Chapter 6.

## 5.1 SIMPLE INDUCTION

Imagine a case in which one is concerned to decide whether a causal generalization found in a base population (say, laboratory mice) also holds true of a target population of interest (say, humans). Simple induction proposes the following rule for such cases:

> Assume that the causal generalization true of the base population also holds approximately in related populations, unless there is some specific reason to think otherwise.

In other words, simple induction proposes that extrapolation be treated as a default inference among populations that are related in some appropriate sense. The advantage of simple induction is that it can be employed in cases in which relatively little detailed information concerning the mechanisms underlying the causal relationship is available. There are, however, three aspects of the above characterization of simple induction that stand in obvious need of further clarification. In particular, to apply the above rule in any concrete case, one needs to decide what it is for a causal generalization to hold approximately, to distinguish related from unrelated populations, and to know what counts as a reason to think that the extrapolation would not be appropriate. It seems doubtful that a great deal can be said about these three issues in the abstract—the indicators of related populations, for instance, can be expected to be rather domain-specific. But it is possible to give examples of the sorts of considerations that may come into play.

Simple induction does not enjoin one to infer that a causal relationship in one population is a precise guide to that in another—it only licenses the conclusion that the relationship in the related target population is



"approximately" the same as that in the base population. It is easy to see that some qualification of this sort is needed if simple induction is to be reasonable. In biology and social science, it is rare that a causal effect in one population is *exactly* replicated even in very closely related populations, since the probabilities in question are sensitive to changes in background conditions. Nevertheless, it is not rare that various qualitative features of a causal effect, such as positive relevance, are shared across a wide range of populations. For example, tobacco smoke is a carcinogen among many human and nonhuman mammal populations. Other qualitative features of a causal effect may also be widely shared; for instance, a fertilizer may promote growth in moderate dosages and inhibit growth in large ones across a wide variety of plant species even though the precise effect differs from one species and variety to the next. In other cases, the approximate similarity may also refer to quantitative features of the causal effect—the quantitative increase in the chance of lung cancer resulting from smoking in one population may be a reasonably good indicator of that in other closely related populations. In the case of extrapolation from animal models, it is common to take into account scaling effects due to differences in body size, since one would expect that a larger dose would be required to achieve the same effect in a larger organism (cf. Watanabe et al. 1992). Thus, in such cases, the scaling adjustment would constitute part of what is covered by "approximately." Depending on the context, the term "approximate" could refer to similarity with regard to any one of the aspects of the causal effect mentioned above, or other aspects, or any combination of them.

Simple induction is also restricted in allowing extrapolations only among related populations, a qualification without which the rule would obviously be unreasonable: no population can serve as a guide for every other. In biology, phylogenetic relationships are often used as a guide to relatedness for purposes of extrapolation: the more recent a shared common ancestor, the more closely related the two species are (cf. Calabrese 1991, 203–4). A phylogenetic standard of relatedness also suggests some examples of what might count as a specific reason to think that the base population is not a reliable guide for the target population. From the mechanistic point of view, phylogenetic relatedness supports extrapolation because it increases the likelihood that the pertinent mechanisms are shared in the base and target populations as the result of descent from a common ancestor. But when the causal relationship in the base population depends on derived features—that is, characteristics not inherited from the common ancestor—this reasoning is fallacious.

In many biological examples, simple induction requires only some relatively minimal background knowledge concerning the phylogenetic relationships among the base and target populations, and its chief advantage lies in this frugality of information demanded for extrapolation. Yet the weakness of the simple inductive strategy also lies in exactly this frugality: given the rough criteria of relatedness, the strategy will inevitably

produce many mistaken extrapolations. According to one review of results concerning interspecies comparisons of carcinogenic effects:

> Based on the experimental evidence from the CPDB [Carcinogenic Potency Database] involving prediction from rats to mice, from mice to rats, from rats or mice to hamsters, and from humans to rats and humans to mice,... one cannot assume that if a chemical induces tumors at a given site in one species it will also be positive and induce tumors at the same site in a second species; the likelihood is at most 49%. (Gold et al. 1992, 583)

A related challenge for simple induction is that it is not rare that there are significant differences across distinct model organisms. For instance, aflatoxin $B_1$ (discussed in section 5.3.2) causes liver cancer in rats but has little carcinogenic effect in mice (Gold et al. 1992, 581–82; Hengstler et al. 2003, 491).

The consequence of these considerations is not that simple induction is wrong or useless for extrapolation. Rather, what follows is that simple induction is limited, and that it is highly desirable that it be supplemented with some more sophisticated inferential strategy. Let us turn to the question of just what this "something more" should be.

## 5.2 POWERS AND CAPACITIES

The notion of a causal power or capacity is a very commonsensical one. For example, in virtue of its hardness and mass, a brick has the capacity to shatter a glass window. Moreover, this capacity is not tied to a specific set of background conditions, but is something that the brick can be reasonably be expected to possess in whatever circumstance it is likely to be found. Capacities and causal powers, then, seem like a promising point of departure from which to address extrapolation. For example, Cartwright maintains that it is only knowledge of capacities that enables one to extrapolate context-dependent relationships such as causal effects from one population to another (1989, 157–58, 163; 1992, 56). According to Cartwright, a statement about a capacity tells us what would occur when all other causes are absent (cf. 1992, 49; 1999, 82–83). But it tells us more than just that, since a capacity exerts its characteristic influence upon the effect even when other causes are present (ibid.). In this section, I argue that capacities approaches to extrapolation have failed to overcome the limitations of simple induction.

The central feature of capacities is their stability across changes in background conditions. As Cartwright puts it, "A property carries its capacities with it, from situation to situation" (1989, 146). This stability need not be absolute (cf. Cartwright 1989, 163), but it is presumably required to be sufficiently robust to justify the expectation that causal influence will hold throughout the domain in question. Capacities are not limited to basic physical properties, such as the mass of a brick. Cartwright also uses the term "capacity" to refer to causal relationships



that depend on a complex set of interactions. For instance, "aspirin's capacity to relieve headaches" is one of her stock illustrations (1989, 141). A claim about the palliative effects of aspirin is quite similar to a statement about the carcinogenic effects of a particular chemical compound. Both are claims about positive causal relevance that depend upon an interaction between a compound and an organism. Thus, the palliative virtues of aspirin exist only in relation to organisms with a particular type of physiology, and similarly for the carcinogenic effects of a particular compound.

Clearly, it will often be difficult to know in advance whether a compound that has a particular effect in one species or class of organisms will have a similar effect in others. That of course is the extrapolation problem of concern in this chapter. But by definition, a capacity is a causal influence that is not tied to a specific context. Hence, if we know *only* that the compound is carcinogenic in (say) rats, we do not know whether its influence can be properly called a capacity. Consequently, if we do not know whether the extrapolation would be correct, we do not know whether the causal effect in question is a capacity. The difficulty here, then, it is that it is not clear how one is to know that something is a capacity independently of already knowing what one wanted to know about extrapolation. In other words, to call a causal relationship a capacity is to say that it is stable across a range of contexts of interest, but questions of extrapolation arise exactly in those cases in which the stability of the causal relationship is in doubt.[2]

The objection that it is unclear how one is supposed to know whether a causal relationship is a stable capacity or merely a local, context-dependent effect has been raised by several authors (Morrison 1995, 165–66; Glennan 1997, 611–13). In response to such concerns, Cartwright writes:

> I have claimed that in the central uses of the concept, we assume that within the specified domain tendencies when properly triggered always "contribute" their characteristic behaviours unless there is a reason why not. (1995, 180)

This statement amounts to a commitment to the use of simple induction: within some set of related populations (the domain), one assumes that the relationship holds unless there is some reason to suppose otherwise. However, we have seen that simple induction is often highly problematic in the context of extrapolation from animal models. Hence, without some further elaboration, the capacities approach will not suffice as a normative account of extrapolation.

Cartwright does provide some elaboration on the issue of whether there is a reason why the capacity will not operate in the new context. This judgment is said to be based upon knowledge of "how this tendency naturally operates and how its power to do so is transmitted, what could distort it, what enhance it, what could damp it and in what ways" (ibid.). This appears to be a reference to Cartwright's notion of a nomological

machine, which is one of several related mechanism concepts, as was discussed in section 3.4.1. On this proposal, then, capacities inhere in the component parts of a nomological machine or mechanism, while extrapolation depends upon information about how the component parts are arranged and interact. That is very similar to the suggestion that knowledge of mechanisms and interfering factors is a basis for extrapolation. But merely to invoke mechanisms is not to have explained how the challenges confront extrapolation. For all we know, the causal effects of the components of the mechanism might be context-dependent, and the components in the model might be arranged and interact differently than those in the target (Alexandrova 2006, 186–87). Demonstrating the relevant similarity of the model and the target would presumably require separately studying the mechanisms in both and then comparing results. But it is not clear how that can be done when the ability to study the target directly is severely limited. In short, to gesture toward mechanisms is not to have answered the challenges confronting extrapolation.

Cartwright's proposal is not unique in this regard: the same point can be made in the context of an account of causal powers provided by Patricia Cheng (1997). Although the aim of Cheng's approach is primarily the psychological one of understanding how people actually draw causal inferences, her proposal is highly interesting from a philosophical perspective. Like Cartwright, Cheng stresses the value of causal powers with regard to extrapolating causal conclusions.

> In the reasoner's mind, causal powers are invariant properties of relations that allow the prediction of the consequences of actions regardless of the causes of an effect (those other than the candidate causes) that happen to occur in a situation. (2000, 127)

Thus, Cheng's causal powers are very similar to Cartwright's capacities in that they are intended to be stable influences that operate independently of changes in context or background conditions. In its simplest version, Cheng's proposal assumes the existence of two types of causes, generative and preventive, which may be either present or absent (but not vary otherwise). No event occurs unless it is caused, and causes can influence their effects only when they are present. An event occurs if and only if at least one of its potential causes is present and causes it on that occasion. For a generative cause C, the causal power of C with respect to E, which we may denote by $p_{ce}$, is the probability that C causes E provided that C occurs. It need not be the case that $p_{ce} = P(E|C)$, since $P(E|C)$ depends not only upon the efficacy of C but also upon the probability of the presence and effectiveness of other causes of E.

Cheng's innovation is to demonstrate that, given certain assumptions, causal powers can be estimated from statistical data (1997, 373–74). In her 1997 paper, one of these assumptions is that $p_{ce}$ is independent of the occurrence of all other causes of E. This means that C, in affecting E, does not interact with any other causes. But in biology and social science,



causes typically influence their effects interactively, so that the impact of one depends upon the presence or absence of others. Indeed, extrapolation is difficult precisely because the relationship between cause and effect might depend on some unknown, variable factor. In light of this limitation, Cheng (2000) develops a concept of interactive causal power. She points out that when causes interact, the formula described in her original proposal does not estimate causal powers, but only what she terms the "contextual causal power" (2000, 235). Cheng also specifies conditions in which interactive causal powers can be estimated from statistical data, provided that *all* the interacting causes have been measured (2000, 241–46). However, in most interesting biological and social science examples, it can be expected that the causes under investigation interact with other causes that have not been measured or otherwise explicitly taken into account. For such cases, Cheng suggests that one proceed by first assuming that the causal power is simple (that is, independent of all other causes) and then postulate causal interactions only when necessary to accommodate conflicting data (2000, 232, 238). Yet the proposal that one estimate context- or population-sensitive causal relationships, and then assume that these hold approximately in related populations unless there is some evidence to the contrary, is simple induction. And as explained in the foregoing section, simple induction is often not a sufficient basis for extrapolation from animal models. Thus, the proposals considered in this section have not provided an adequate account of how extrapolation could proceed even when not justifiable by simple induction. Let us turn, then, to a distinct proposal.

## 5.3 MECHANISMS-BASED EXTRAPOLATION

The mechanisms approach to extrapolation suggests that knowledge of mechanisms and factors capable of interfering with them can provide a basis for extrapolation. But this proposal must also answer the two challenges to extrapolation described above. Since causally relevant differences between model and target are inevitable, some explanation must be provided of how extrapolation can be justified even when there are some differences in mechanism between model and target. The extrapolator's circle confronts the mechanisms proposal as well. Presumably, justifying the appropriateness of the model would involve comparing mechanisms in the model and the target, which would involve independently studying the mechanisms in both and then comparing results. But that makes it unclear how the suitability of the model can be established without already knowing what the extrapolation was supposed to tell us. In this section, I argue that existing discussions of mechanisms do not adequately address these challenges. Then I present a more adequate account of mechanisms-based extrapolation that is founded upon what we call *comparative process tracing*.

### 5.3.1 The Existing Literature on Mechanisms and Extrapolation

There is a small literature that provides detailed case studies of extrapolation in biology or social science, often with particular attention to the role of mechanisms (cf. Burian 1993; Ankeny 2001; Schaffner 2001; Weber 2005; Guala 2005; Alexandrova 2006). Essays in this genre point out some circumstances that facilitate, and some that hinder, extrapolation. For instance, it has been observed that extrapolation is on firmer ground with respect to basic, highly conserved biological mechanisms (Wimsatt 1998; Schaffner 2001; Weber 2005, 180–84). Others have observed that a close phylogenetic relationship is not necessary for extrapolation and that the use of a particular animal model for extrapolation must be supported by empirical evidence (Burian 1993). Similarly, Francesco Guala (2005) emphasizes the importance in experimental economics of providing empirical evidence to support the claim that the model is relevantly similar to the target.

These suggestions are quite sensible. The belief that some fundamental biological mechanisms are very widely conserved is no doubt a motivating premise underlying work on such simple model organisms as the nematode worm. And it is certainly correct that the appropriateness of a model organism for its intended purpose is not something that may merely be assumed, but a claim that requires empirical support. Yet such observations do not answer the challenges to extrapolation. Objections to animal extrapolation focus on causal processes that do not fall into the category of fundamental, conserved biological mechanisms. For example, Marcel Weber suggests that mechanisms be conceived of as embodying a hierarchical structure, wherein the components of a higher-level mechanism consist of lower-level mechanisms, and that while lower-level mechanisms are often highly conserved, the same is not true of the higher-level mechanisms formed from them (2001, 242–43; 2005, 184–86). So, even if one agreed that basic mechanisms are highly conserved, this would do little to justify extrapolations from mice, rats, and monkeys to humans regarding such matters as the safety of a new drug or the effectiveness of a vaccine. Since critiques of animal extrapolation are often motivated by ethical concerns about experimentation on animals capable of suffering (cf. LaFollette and Shanks 1996), they primarily concern animal research regarding less fundamental mechanisms that cannot be studied in simpler organisms such as nematode worms or slime molds. Nor do the observations sketched in the foregoing paragraph explain how extrapolation can proceed even when there are causally relevant differences between model and target or how the extrapolator's circle is to be avoided. For example, noting that the appropriateness of an animal model for a particular extrapolation is an empirical hypothesis does not explain how such a hypothesis can be established without already knowing what one wishes to extrapolate.

There also are discussions in the philosophical literature of strategies for learning about mechanisms. A distinction between mechanisms and



the "phenomena" (Craver and Darden 2001, 113–14) or "behavioral descriptions" (Glennan 2005, 446) those mechanisms explain is helpful for understanding these proposals and contrasting them with comparative process tracing. Phenomena are regularities of the system under study that are more easily observable than the underlying mechanisms. For example, that HIV exposure causes AIDS is a phenomenon, whereas the mechanism consists of the molecular processes through which HIV has this effect. Since phenomena are often more easily discovered than underlying mechanisms, several authors have examined strategies for discovering mechanisms, given the phenomenon and some background constraints on what the components of the mechanism and their interactions could be (cf. Bechtel and Richardson 1993; Craver and Darden 2001; Darden and Craver 2002). Lindley Darden and Carl Craver's (2002) discussion focuses on what they term *schema instantiation* and *forward chaining/backtracking*. Schema instantiation begins with a schematic outline of the mechanism in which central functional roles are specified, but important details concerning the entities and activities involved in the performance of those functions are omitted. For example, the mechanism of HIV replication instantiates a schema that is common for retroviruses: attachment to a target cell, insertion of viral RNA into cytoplasm, reverse transcription, integration of viral DNA into host DNA, and synthesis of products for the formation of new viruses from this integrated viral DNA by means of the host cell's genetic machinery. Next, one attempts to discover the specific entities and activities that instantiate the schema, often by means of tracing forward from a known starting point or backward from a known end point (or both at once). For convenience, we will refer to the joint application of these strategies, schema instantiation and forward chaining/backtracking, as *process tracing*.

Glennan points out that process tracing is sometimes unfeasible for ethical or practical reasons (2005, 459–61). In such cases, one may attempt to discover the mechanism through more detailed descriptions of the phenomenon (ibid.). For example, alternative hypotheses concerning the mechanism may yield differing predictions about how the system would behave in a new circumstance. But although the strategy that Glennan suggests differs from process tracing, it is aimed at solving the same inference problem: *given* a description of the phenomenon, *discover* the mechanism that accounts for it. In extrapolation, by contrast, what one wishes to infer is a mechanism and phenomenon in a target organism. The evidence given includes the mechanism and behavioral description for a model organism, and perhaps some partial information about the mechanism in the target. By "partial information," I mean that the information concerning the mechanism in the target is not sufficient on its own to infer the phenomenon (e.g., whether the compound is carcinogenic in humans). The mechanisms approach to extrapolation must indicate a strategy for solving the following inference problem: *given* both the mechanism and the phenomenon *in the model*, and partial information

concerning the mechanism in the target, *infer* the mechanism and/or phenomenon *in the target*.

### 5.3.2 Comparative Process Tracing

Suppose that one is given a description of the mechanism in the model organism and wishes to use this information as a basis for extrapolation. Such an inference is a case of reasoning by analogy. The form of arguments by analogy can be represented schematically as follows: the base (or source or analogue) is known to possess properties 1 through $n$, while the target is known to have properties 1 through $n-1$; therefore, the target also possesses property $n$. It is obvious that not all inferences satisfying this abstract schema are reliable. For instance, Bob and Sue may both own 2005 Volkswagen Beetles, yet the information that Bob's car is iridescent lime green provides little support for the conclusion that Sue's car is the same color (cf. Weitzenfeld 1984, 138; Davies 1988, 229). Arguments instantiating the above schema, then, provide substantial support for their conclusions only given some additional, perhaps implicit, information. This additional information would consist of generalizations asserting that objects of specified types typically resemble one another in certain ways, though not necessarily in others. For instance, suppose one wanted to know whether the engine in Sue's Volkswagen Beetle is in the rear of the car (as in the older models) or in the front. If we learned that the engine of Bob's car is front mounted, we readily conclude that the same is true of Sue's car. The difference between this analogical inference and the one above is that cars of the same make, model, and year are typically manufactured in a variety of colors yet are generally similar with regard to basic design features such as the placement of the engine. Likewise, mechanisms-based extrapolation depends on knowledge of likely similarities and dissimilarities of the mechanisms between model and target.

If one peruses a text or review article on animal extrapolation in toxicology, one finds a compendium of information concerning how pertinent mechanisms differ between humans and various model organisms, and with respect to which types of compounds.[3] In the case of carcinogenesis, probably the most frequent differences concern metabolism (Calabrese 1991, chap. 5; Hengstler et al. 1999, 918). Since the metabolism of foreign, potentially toxic compounds consists of chemically transforming them so as to make them less toxic and more readily excreted, differences with regard to how a particular compound is metabolized, and at what rate, can have implications for its carcinogenic effects. Mechanisms for metabolism of foreign compounds are typically described in terms of two phases (cf. Calabrese 1991, 206). In phase I, the compound is chemically altered (often through the addition of oxygen or hydrogen atoms) in a manner that makes it more polarized, and consequently more easily excreted. In phase II, the compound resulting from the modification in phase I is conjoined with a macromolecule, such as a carbohydrate, which typically



detoxifies the compound and further facilitates its removal. Metabolic mechanisms can differ with respect to how the compound is altered at either phase and in virtue of which enzymes catalyze the process, which has the result that some mechanisms may be more effective than others at detoxifying and eliminating a given foreign compound.

The above discussion suggests a procedure for extrapolating a mechanism found in the base population to the target population, a procedure that I call *comparative process tracing*. First, learn the mechanism in the model organism, by means of process tracing or other experimental means. For example, a description of a carcinogenic mechanism would indicate such things as the product of the phase I metabolism and the enzymes involved; whether the metabolite is a mutagen, an indication of how it alters DNA; and so on. Second, compare stages of the mechanism in the model organism with that of the target organism in which the two are most likely to differ significantly. For example, one would want to know whether the chemical is metabolized by the same enzymes in the two species, and whether the same metabolite results, and so forth. In general, the greater the similarity of configuration and behavior of entities involved in the mechanism at these key stages, the stronger the basis for the extrapolation.

The reliability of comparative process tracing depends on correctly identifying the points at which significant differences between the model and the target are likely to arise. Significant differences are those that would make a difference to whether the causal generalization to be extrapolated is true in the target. For instance, metabolism is a source of potentially significant difference in carcinogenesis, since how a compound is metabolized often matters to whether it is carcinogenic or not. Judgments about where significant differences are and are not likely to occur are based on inductive inferences concerning known similarities and differences in related mechanisms in a class of organisms, and on the impact those differences make. In the present case, the relevant generalizations would concern the common similarities and significant differences in carcinogenic mechanisms between humans and rodents. Comparative process tracing, then, resembles simple induction in relying upon generalizations concerning the relation between the target and model organisms. The chief difference concerns what these generalizations assert. Simple induction depends upon generalizations of the form "What is carcinogenic for rats is probably carcinogenic for humans, too." In contrast, comparative process tracing depends upon generalizations like "Features *A*, *B*, and *C* of carcinogenic mechanisms in rodents usually resemble those in humans, while features *X*, *Y*, and *Z* often differ significantly." The toxicology literature described above is plausibly interpreted as an effort to provide an empirical basis for generalizations of the latter but *not* the former sort. Of course, it might be questioned whether the data presently available to toxicologists constitute a representative sample. However, that is a standard problem of statistical sampling rather than

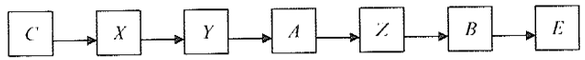



Figure 5.1 Comparing a downstream stage

a difficulty specifically raised by extrapolation, such as the extrapolator's circle.

But even given accurate information about the points of likely similarity and dissimilarity, comparative process tracing might still be impractical if not all likely points of significant difference could be compared. Fortunately, comparative process tracing often does not require comparing every stage of the mechanism at which significant differences are likely to be present. In particular, suppose that many points of likely difference are upstream of a later stage that is relatively easy to measure and compare. Then it may be possible to omit comparisons of the upstream stages and focus on the downstream one. For instance, imagine a mechanism like the following:

Suppose that *X*, *Y*, and *Z* represent points of the mechanism at which significant differences between model and target are likely, while *A* and *B* represent points that are likely to be the same. If differences in *X* or *Y* must result in differences in *Z*, then it is necessary only to compare the model and target at *Z*. That reduces the amount of information about the mechanism in the target that is needed to establish the suitability of the model, which may be very helpful if it is difficult to study the mechanism in the target directly. Furthermore, comparing a downstream stage of the mechanism also renders mistakes about upstream sources of difference less consequential. For instance, suppose that differences were in fact likely at *A* in Figure 5.1, despite our belief to the contrary. Yet if a difference at *A* must generate differences at *Z*, then the mistaken belief about *A* will not lead to a faulty extrapolation so long as a comparison is made at *Z*. Thus, efficient applications of comparative process tracing can focus on likely sources of difference in *downstream* stages of the mechanism.

A few important qualifications about the emphasis on downstream stages should be noted. First, the strategy could lead to mistaken conclusions if there is a path that bypasses the downstream stage. For instance, suppose in Figure 5.1 there was a path from *X* to *E* that did not go through *Z*. In that case, checking *Z* would not be sufficient since there might be significant differences in the mechanisms that would not leave a mark on *Z*. Hence, applications of the strategy depend on knowing where to look for bottlenecks through which any influence upon the outcome must be transmitted. Second, the mark that upstream stages leave upon the downstream stages must be distinctive in the sense that it could not have resulted from some independent cause. The mark should be, as it were, a fingerprint whose presence or absence indicates something causally significant about upstream processes. In examples from toxicology, the distinctive mark is often a particular chemical compound that retains



a distinctive, identifiable structure even after being metabolized. That point is illustrated by the aflatoxin $B_1$($AFB_1$) example that we discuss now.

Extrapolation of the carcinogenic effects of aflatoxin $B_1$($AFB_1$) is a good example of comparative process tracing. Produced by certain species of fungi that grow on various types of grains and nuts, aflatoxins are now generally regarded as an important risk factor for liver cancer, a belief dating back to the 1960s that has its origins in laboratory experiments on rats and epidemiological studies (Wogan 1992, 123). Jointly, a positive correlation in epidemiological data between liver cancer and exposure to aflatoxins through food contamination, and a corresponding experimental result in rats, provided a prima facie case for the conclusion that aflatoxins are carcinogenic in humans. However, this evidence alone is not unequivocal. The epidemiological correlation might result in whole or in part from an unmeasured common cause of aflatoxin exposure and liver cancer, while rats might be an inappropriate model for humans with respect to aflatoxins. The appropriateness of the rat as a model in this context was hardly an idle concern, given that aflatoxin was found to have little carcinogenic effect in mice (Gold et al. 1992, 581–82; Hengstler et al. 2003, 491). Differing results among animal models are a clear case of a "reason to suppose otherwise," blocking extrapolation by simple induction. Let us consider how comparative process tracing ameliorated this situation.

Since there are often trans-species differences in the metabolism of foreign compounds, a natural starting point for this inquiry was to analyze the metabolism of aflatoxins in humans and in the rodent populations in which aflatoxins were found to be carcinogenic. It was found that $AFB_1$, the most common aflatoxin, was converted to the same phase I metabolite across these groups (Wogan 1992, 124). Given the sharp differences in carcinogenic effects of $AFB_1$ in rats and mice, it was of obvious interest to inquire which of these two animal models was a better guide for humans. It was found that although the phase I metabolism of $AFB_1$ proceeded similarly among mice, rats, and humans (and in fact at a higher rate in mice), the phase II metabolism among mice was extremely effective in detoxifying $AFB_1$ but not among rats or humans (Hengstler et al. 1999, 928–31). Furthermore, this metabolite bound to DNA in rat liver cells in vivo at sites at which the nucleotide base guanine was present to form complexes called DNA adducts (ibid., 927). It was further found that such cells suffered unusually frequent mutations in which guanine-cytosine base pairs were replaced with adenine-thymine pairs, a mutagenic effect found in vivo among rats and in vitro among cells of a variety of origins, including bacteria and human (ibid., 923, 927). In addition, guanine-cytosine to adenine-thymine mutations were found in activated oncogenes present in rats exposed to $AFB_1$ but were absent in the controls (ibid., 130–33). Thus, comparative process tracing yielded the conclusion that the rat was a better model than the mouse

The example also illustrates that comparative process tracing need not be restricted to comparisons between a single model-target pair, but may involve selecting among several candidate model organisms. In fact, rats and mice were not the only model organisms considered: guinea pigs and hamsters were also studied. These were compared with humans on the basis of quantity of $AFB_1$ DNA adducts present per unit of peripheral blood among individuals exposed to $AFB_1$, with a one strain of rat, the Fischer rat, bearing the closest similarity to humans (Hengstler et al. 1999, 925–26). However, even in the Fischer rat, the quantity of DNA adducts was significantly less than in humans, suggesting that even the most sensitive rodent model provides an underestimate of the human impact of $AFB_1$. The quantity of DNA adducts provides information about a downstream stage of the mechanism (like Z in Figure 5.1). Thus, by focusing on the quantity of DNA adducts, researchers could avoid the cumbersome task of comparing every likely point of difference. This example also demonstrates how comparative process tracing can indicate extrapolative limitations of the best model. In this case, one could reasonably use the Fischer rat to extrapolate the conclusion that $AFB_1$ exposure increases the chance of liver cancer, and perhaps even use the effect in the Fischer rat to estimate a lower bound for the strength of that effect. But it is doubtful that a quantitative estimate of the impact of $AFB_1$ upon liver cancer could be correctly extrapolated from the Fischer rat to humans.

## 5.4 CRITIQUES OF ANIMAL EXTRAPOLATION

An account of extrapolation should be able to adjudicate methodological disputes on this topic, and this section illustrates how the proposal advanced here can do that. In a book and series of articles, Hugh LaFollette and Niall Shanks argue that model organisms cannot be reliably used for extrapolation at all, but only as sources of promising hypotheses to be tested by clinical or epidemiological investigations (1993a, 1993b, 1995, 1996). They use the term *causal analogue model* (CAM) to refer to models that can ground extrapolation, and *hypothetical analogue model* (HAM) to refer to those that function only as sources of new hypotheses. According to LaFollette and Shanks, animal models can be HAMs but not CAMs. A similar though somewhat more moderate thesis is advanced by Weber. He maintains that, except for studies of highly conserved mechanisms, animal models primarily support only "preparative experimentation" and not extrapolation (2005, 185–86). Weber's "preparative experimentation" is similar to LaFollette and Shanks's notion of a HAM, except that it emphasizes the useful research materials and procedures derived from the animal model in addition to hypotheses (2005, 174–76, 182–83). In this section, I argue that these pessimistic claims about the potential of animal extrapolation are not correct.



### 5.4.1 No Relevant Difference

LaFollette and Shanks's primary argument for the conclusion that model organisms can function only as HAMs and not as CAMs rests on the proposition that if a model is a CAM, then "*there must be no causally relevant disanalogies between the model and the thing being modeled*" (1995, 147; italics in original).[4] It is not difficult to show that animal models rarely if ever meet this stringent requirement. But an obvious reply is that LaFollette and Shanks's criterion of CAM-hood is unreasonably strict. In light of this, LaFollette and Shanks consider the possibility that a weaker condition than the complete absence of relevant causal disanalogies could suffice for extrapolation. They suggest that this proposal be interpreted as follows:

> Begin with two systems, $S_1$ and $S_2$. $S_1$ has causal mechanisms [a, b, c, d, e], $S_2$ has mechanisms [a, b, c, x, y]. When stimulus $s_f$ is applied to subsystems [a, b, c] of $S_1$, response $r_f$ regularly occurs. We can therefore infer that were $s_f$ applied to subsystems [a, b, c] of $S_2$, it is highly probable that $r_f$ would occur. (1995, 153)

However, they argue that this inference is valid only if the relationship between the stimulus and the response is entirely independent of the differing mechanisms, [d, e] and [x, y] (ibid.). But if these mechanisms make no difference to the relationship between the stimulus and the response, then there are no relevant disanalogies between $S_1$ and $S_2$, which would mean that $S_1$ is a CAM after all. Thus, LaFollette and Shanks conclude that when it comes to extrapolation, only a CAM in their sense will do: there must be no relevant causal dissimilarities between model and target (cf. 1996, 180).

Needless to say, this strict condition is rarely if ever satisfied. Not only are relevant differences across species inevitable, but dissimilarities are also extremely common *within species* and even for a *single organism* at different stages of its life. The field of pharmacogenomics, for instance, is dedicated to the study of genetic differences among humans that produce divergent responses to drug therapies. Likewise, susceptibility to, say, harmful side effects of a therapy may be contingent upon factors associated with age, such as declining kidney functioning. Thus, if the strict criterion of CAM-hood proposed by LaFollette and Shanks were accepted, not only would extrapolation from animal to human be illegitimate, but so would extrapolation from humans to other humans. Indeed, even extrapolations from past to future in the life of a single person would be unjustified.[5]

The flaw in LaFollette and Shanks's argument is that it overlooks the connection between the specificity of the claim to be extrapolated and the standard of a suitable model. This point is illustrated nicely by the aflatoxin example. In this case, the Fischer rat would not qualify as a CAM in LaFollette and Shanks's strict sense, since the quantity of DNA adducts

resulting from $AFB_1$ is less in the Fischer rat than in humans. Yet this difference does not undermine extrapolating the positive causal relevance of $AFB_1$ for liver cancer. The difference suggests that the effect in the Fischer rat is *less* than that in humans. But if the effect in Fischer rats is positive and less than that in humans, then the effect in humans must be positive, too. Consequently, although it would be unwise to extrapolate the *exact* causal effect of $AFB_1$ upon liver cancer from Fischer rats to humans, the known difference provides no reason against extrapolating a claim about positive causal relevance. Thus, a model might provide a good basis for extrapolating a *qualitative*, but *not a quantitative*, claim concerning a causal effect.

This example suggests that LaFollette and Shanks's stringent criterion of CAM-hood is simply a characterization of what a model organism must be if it is to serve as a basis for the extrapolation of *exact* causal effects. Generally, neither animal-model-to-human nor human-to-human extrapolation can expect such precision. For instance, there is reason to think that the quantitative effect of $AFB_1$ upon liver cancer varies among human populations. One important reason is that exposure to the hepatitis B virus appears to increase susceptibility to the carcinogenic effects of $AFB_1$ (cf. Kew 2003), and rates of exposure to that virus vary geographically. LaFollette and Shanks's mistake, therefore, is to present their characterization of a CAM as an entirely general condition required for the extrapolation of any causal claim whatever, when it is in fact only a criterion for extrapolating an extremely precise causal generalization. The conditions that suffice for extrapolating claims concerning positive causal relevance are far less stringent than those needed for extrapolating the exact probability distribution of the effect, conditional on interventions that set the value of the cause.

Chapter 6 explores in greater generality and precision conditions that suffice for extrapolating claims concerning positive or negative causal relevance. In section 6.2.2, I explain how these sufficient conditions are in fact quite reasonable in the aflatoxin example.

### 5.4.2 The Extrapolator's Circle

LaFollette and Shanks also use the extrapolator's circle as an argument for their conclusion that animal models can function only as HAMs and not as CAMs. They claim, reasonably enough, that the appropriateness of a model organism for extrapolation must be demonstrated by empirical evidence (1993a, 120).[6] But they argue that this appropriateness cannot be established without already knowing what one hopes to learn from the extrapolation.

> We have reason to believe that they [animal model and human] are causally similar only to the extent that we have detailed knowledge of the condition in *both* humans and animals. However, once we have enough information to be confident that the non-human animals are causally similar (and thus, that inferences from one to the other are probable), we likely know most of what the CAM is supposed to reveal. (1995, 157)[7]



LaFollette and Shanks presumably mean to refer to their strict CAM criterion when they write "causally similar," but the extrapolator's circle can be stated independently of that criterion. Whatever the criterion of a good model, the problem is to show that the model satisfies that criterion given only limited, partial information about the target.

However, LaFollette and Shanks's argument shows that extrapolation from animal to human is never legitimate only if it proves the same for extrapolation from one human group to another. For suppose that a particular causal generalization is known to obtain in one human population, and the question is whether it does so in a second. How is one to know whether the two populations are sufficiently similar for the purposes of the extrapolation? According to LaFollette and Shanks, this similarity can be established only on the basis of independently learning the causal relationship in each population and then comparing results. But that would obviate the need for the extrapolation. Thus, the extrapolator's circle shows that animal extrapolation is never justified only if it shows the same about extrapolation in all heterogeneous populations.

This result suggests that the extrapolator's circle does not really show that animal extrapolation can never justify informative conclusions about humans. An account of extrapolation should be able to specify where LaFollette and Shanks's argument goes wrong, while indicating the extent to which the extrapolator's circle is a genuine problem. Unlike previous accounts of extrapolation, the proposal advanced here can do that. LaFollette and Shanks's attempt to turn the extrapolator's circle into a general critique of animal extrapolation overlooks the role of premises concerning likely similarities and differences in analogical reasoning. Thus, in comparative process tracing, providing evidence for the suitability of the model requires comparisons *only* at stages in the mechanism in which significant differences are likely to occur. Consequently, it may be necessary to compare only a few stages of the mechanism. For example, metabolism is the most common source of difference in carcinogenic mechanisms among mammals. Thus, showing that phase I and II metabolism of $AFB_1$ proceeds similarly in rats and humans strengthens the case for the rat as a model organism. Yet an understanding of the phase I and II metabolism of $AFB_1$ in humans, considered on its own, provides little information regarding the carcinogenic effects of this compound. Moreover, it is not necessary to compare all points of likely significant difference if there is a downstream stage of the mechanism upon which upstream differences leave their mark. This point is illustrated in the $AFB_1$ case by the use of the quantity of DNA adducts to assess several potential animal models. In sum, making a case for the suitability of the model may require examining only a few key features of the mechanism in the target, and knowledge of these features alone would fall far short of what one hopes to learn from the extrapolation. In such cases, the extrapolator's circle is avoided.

The extrapolator's circle is a serious challenge if little is known about likely similarities and differences in relevant mechanisms or if it is known that the model and the target are likely to differ in almost every relevant respect. In the latter case, one would effectively know that the organism in question is in fact a very poor model, which would imply that it ought not to be used as a basis for extrapolation. The more interesting case, then, is the first: little is known about likely similarities and differences or their significance for the causal relationship in question. There can be little doubt that such cases sometimes arise, and when they do, extrapolation obviously cannot proceed by comparative process tracing, but would presumably rely upon simple induction. But transforming the extrapolator's circle into a general critique of extrapolation from animal models would require not merely showing that such circumstances *sometimes* arise. It would be necessary to show that this situation is *almost always* the one faced in animal extrapolation. That is an argument that LaFollette and Shanks have not made, and it is one that seems difficult to make, given examples like aflatoxin.

That comparative process tracing can establish the suitability of an animal model also demonstrates that extrapolation is not restricted to entrenched mechanisms inherited from distant ancestors. The carcinogenic mechanism in the $AFB_1$ example is clearly not of this character since, for instance, it is not present in mice. In short, that a mechanism is highly conserved is *one, but not the only,* possible basis for extrapolation.

### 5.4.3 HAM Versus CAM?

An underlying assumption of LaFollette and Shanks's argument is that there is a sharp divide between CAMs, which can support extrapolation, and HAMs, which only suggest fruitful hypotheses and lines of research. They write that "there is a big difference between an animal model being a good source of hypotheses and its being a good means to test hypotheses" (1996, 199). LaFollette and Shanks support the claim that there is a strict divide between HAM and CAM by appeal to the old distinction between the contexts of discovery and justification (1996, 194).[8] According to this doctrine, the manner by which a hypothesis is generated has no relevance whatever to the assessment of its scientific adequacy. Whether the new hypothesis was inspired by a dream, a poem, or the floral pattern of a colleague's Hawaiian shirt makes no difference to its epistemic virtues, which can be decided only through a careful examination of the relevant evidence. The sharp contrast between HAM and CAM drawn by LaFollette and Shanks is simply the context of discovery versus justification distinction applied to animal models. HAMs are animal models in the context of discovery, while CAMs are models in the context of justification.

However, the context of discovery versus justification dichotomy has been critiqued from a wide variety of perspectives (cf. Hanson 1958; Kuhn 1977, chap. 11; Longino 1990; Darden 1991; Kelly 1996; Simon 1998).



Current discussions of the distinction in the philosophy of science literature take it as more or less given that aspects of the discovery process can be relevant to the assessment of hypotheses, and then proceed to consider the finer points of proposals about how this is so (cf. Darden and Craver 2002; Castle 2001; Elliott 2004). The problem with the thesis that there is an unbridgeable chasm between the contexts of discovery and justification can be appreciated by means of simple examples like the following. Imagine two procedures for generating hypotheses, the first of which generates correct hypotheses 95 percent of the time and the second that generates correct hypotheses 1 percent of the time. Now suppose that the two procedures have produced conflicting hypotheses. Given this information, which hypothesis—the one generated by the first procedure or the one generated by the second—do you think is more likely to be correct?

The obvious answer is the hypothesis produced by the first procedure. Thus, that a hypothesis was generated by a procedure likely to produce empirically successful hypotheses can be relevant evidence. Although it is rarely possible to assign exact rates of success to distinct discovery procedures, the process is nevertheless typically not a matter of ineffable and mysterious inspiration either. For example, scientific discovery is typically guided by prior knowledge of constraints that must be satisfied by a successful hypothesis in the domain in question. Ignoring these constraints is likely to lead to a grossly inadequate hypothesis. In sum, the process by which hypotheses are discovered is amenable to logical analysis and can be relevant evidence to be considered in assessing the hypothesis.

A defender of the context of discovery versus justification dichotomy might object that the mode of discovery is evidentially relevant only insofar as it suggests that the hypothesis is consistent with particular observations or experimental results. Consequently, the mode of discovery would be irrelevant to one who knew all of these data and who was able to directly assess the hypothesis with regard to them. That may be true, but it is nevertheless the case that information about the mode of discovery may be evidentially relevant to someone in a less than perfect epistemic position. One might not know what all of the relevant data are, or one might not be able to directly assess whether the hypothesis is consistent with them. In such cases, information regarding the source of the hypothesis may remain evidentially relevant. This is very much the situation one faces with regard to animal extrapolation. For instance, to a person with complete knowledge of carcinogenesis in humans, information about animal models would be irrelevant for assessing the accuracy of any hypothesis about the effects of $AFB_1$. But for ordinary mortals who lack such perfect knowledge, animal models can be a useful source of evidence.

These considerations are directly relevant to the supposed sharp distinction between HAM and CAM. As LaFollette and Shanks observe

although hypotheses can be inspired by practically anything, not everything is a good HAM (1996, 195). The most obvious way a model could be a good HAM is in virtue of being likely to generate hypotheses about the target that are true, or at least approximately so. Yet this account of what makes a good HAM entails that the difference between HAM and CAM is one of degree. Both provide some evidence for the extrapolation; it is just that the evidence provided by the CAM is stronger and less equivocal. But LaFollette and Shanks cannot distinguish between good and bad HAMs in this manner, since that would contradict their claim that *only* CAMs in their very strict sense provide *any* evidence for extrapolation.

So what does make a good HAM, according to LaFollette and Shanks? They write, "A HAM is likely to be valuable if there are demonstrable functional similarities between the model and item modeled" (1996, 195). But it is difficult to see how this could be true, given their persistent claim that functional similarity is no indicator of similarity of mechanisms.[9] For in that case, there is no reason to think that a functionally similar HAM will lead to fruitful hypotheses rather than unproductive dead ends. Of course, model organisms typically share more with their targets than mere functional similarity. They also share a common ancestor and some fundamental mechanisms at the level of biochemistry, the cell, and physiology. These similarities provide some—albeit rather uncertain and rough—grounds for extrapolation. And that is what justifies regarding them as HAMs.

Rather than a sharp dichotomy between HAMs and CAMs, then, there is a continuum from models providing weaker to those providing stronger grounds for extrapolation. A model might be a weak basis for extrapolation because little is known about likely sources of significant difference and similarity, or because mechanisms in the model and the target have not been compared at stages of likely difference. The more that is known about likely similarities and differences, and the more the likely differences have been checked and found to be absent, the stronger the basis for extrapolation. Moreover, exactly how similar the model is required to be depends upon the claim of interest to the extrapolation, as noted above and explored in further detail in Chapter 6. From this perspective, any sharp HAM versus CAM distinction is inevitably arbitrary and ultimately unimportant. The pertinent issues are how thoroughly comparative process tracing has been carried out and what conditions are required to extrapolate the generalization in question.

Despite disagreeing with LaFollette and Shanks's methodological critique of animal extrapolation, I think that they deserve credit for articulating objections that had not been adequately addressed in the literature on this topic. I also think that they are correct that these methodological questions matter to ethical issues surrounding animal experimentation, since animal research is typically justified on the grounds that it provides knowledge that benefits humans. Thus, the ethical question turns on whether the benefit to humans outweighs the suffering of the animal



model. Although an in-depth exploration of these ethical issues is beyond the scope of this book, I would like to briefly indicate what I regard as the main ethical implication of the account of extrapolation developed here. LaFollette and Shanks wish to argue that animal experimentation is unethical in general, and hence they endeavor to show that extrapolation, in general, is not a reliable source of new information about humans. And if animal extrapolation were indeed so utterly incapable of providing useful information concerning humans, then the standard ethical defense of animal research would be undermined. In contrast, I suggest that extrapolation is reliable and informative in some circumstances but not others, and make some steps toward clarifying what those circumstances are. This perspective calls across-the-board moral vindications or condemnations of animal research into question.[10] Whether animal research is ethically defensible in a given case may depend *in part* upon the potential for extrapolating useful information about humans. And the extent to which this is or is not possible will depend on complex, case-specific scientific details. I do not pretend to answer the question of whether and to what extent animal research is ethically defensible. However, I do think that my account of extrapolation casts doubt on any "one size fits all" argument on either side of the issue.

## 5.5 CONCLUSION

This chapter presents a mechanisms approach that addresses some of the primary methodological challenges confronting animal extrapolation. I began by considering simple induction, which is an undeniably important aspect of extrapolation but also limited in important ways. Simple induction alone would result in many mistaken extrapolations from animals to humans. In addition, there often is some reason to suppose that the extrapolation might be inaccurate, and simple induction provides little guidance about what to do when that is the case. More sophisticated approaches to extrapolation attempt to indicate how the suitability of a model for a particular extrapolation could be established. Any proposal of this sort must confront what I called the extrapolator's circle. That is, it must explain how the suitability of the model could be established without already knowing what the extrapolation is supposed to tell us. Moreover, since causally relevant disanalogies between animal models and human targets are inevitable, it is necessary to explain how extrapolation can be legitimate even when such disanalogies are present. I argued that existing proposals concerning extrapolation—either in terms of capacities or in terms of mechanisms—fail to adequately address either of these challenges. However, I proposed that the mechanisms approach can be developed so as to provide an answer to the extrapolator's circle. The key proposition in this proposal is what I called comparative process tracing. Comparative process tracing depends upon possessing information about the stages at which significant differences in mechanisms are and are not

likely to occur, and on the directional property of the mechanism which enables one to focus on downstream stages when looking for relevant difference. Thus, it may be possible to establish the suitability of a model organism through a comparison with the target at a small number of stages in the mechanism. Finally, I examined several general methodological objections to animal extrapolation that were motivated by concerns about the ethical permissibility of animal research from the perspective of the approach to extrapolation proposed in this chapter. Although I think that these objections raise important issues, I argued that they are unsuccessful. In the next chapter, I explore conditions that can justify extrapolating claims of positive or negative causal relevance in greater detail, and suggest that this topic is closely relevant to the issue of ceteris paribus laws.

# 6

# *Ceteris Paribus* and Extrapolation

Laws and generalizations qualified by the expression "*ceteris paribus*," a Latin phrase for "other things being equal," are argued by some to play an important role in biology and social science. In contrast, others object that there is no satisfactory interpretation of ceteris paribus (hereafter, cp) laws and that they are not useful for understanding characteristic generalizations in the biological or social sciences. This chapter examines the controversy over cp laws from the perspective of extrapolating claims about positive or negative causal relevance. I propose that considering the topic in this light helps to resolve a central puzzle concerning the scientific role of cp laws.

A survey of the current literature on the topic reveals that the expression "cp law" is highly ambiguous: several types of generalizations have been classified under this label. This point has been made explicitly by Gerhard Schurz (2001b, 2002), and it is also implicit in the variety of proposals concerning the manner in which cp laws should be understood. On some of these interpretations, cp laws are in fact illustrated by causal claims encountered in earlier chapters, such as causal effects and descriptions of mechanisms. The issue of cp laws is also related to extrapolation: one might say that a causal generalization found in one context will also obtain in another, provided that *nothing interferes* or *all else being equal*. That is, the expression "ceteris paribus" can serve as a vague, all-purpose term for indicating whatever conditions are needed for the extrapolation to be correct. Moreover, extrapolation is an important part of what motivates discussions of cp laws, since the content of the cp clause is intended to provide guidance about when the generalization can and cannot be appropriately applied.

A common type of analysis of cp laws known as the "completer approach" interprets laws as universal generalizations and the cp clause as stating conditions in which the law holds without exception. But in cases in which the conditions that lead to exceptions to the law cannot be listed exhaustively, the completer approach inevitably violates what I call the *domain specificity criterion*. This criterion requires that a law of a domain should provide information specifically about that domain rather than merely asserting, say, that determinism is true. I propose that the failings of the completer approach arise from two sources. First, it interprets "ceteris paribus" as qualifying a *generalization* in cases in which that expression should be understood in reference to an *inference schema*. Unlike an empirical law, an inference schema (such as *modus tollens*)