Philip Kitcher

Explanatory Unification and the Causal Structure of the World

1. Introduction

The modern study of scientific explanation dates from 1948, the year of the publication of the pioneering article by C. G. Hempel and Paul Oppenheim. Nearly forty years later, philosophers rightly continue to appreciate the accomplishments of the covering-law models of explanation and the classic sequence of papers in which Hempel articulated his view. Even though it has become clear that the Hempelian approach to explanation faces difficulties of a number of types, the main contemporary approaches to explanation attempt to incorporate what they see as Hempelian insights (with distinct facets of the covering-law models being preserved in different cases), and they usually portray themselves as designed to accommodate one or more of the main problems that doomed the older view. My aim in this essay is to compare what I see as the chief contemporary rivals in the theory of explanation, to understand their affiliations to the covering-law models and their efforts to address the troubles of those models, and to evaluate their success in doing so. Ecumenical as this may sound, the reader should be forewarned that I shall also be interested in developing further, and defending, an approach to explanation that I have championed in previous essays (1981, 1985c).

1.1 Hempel's Accounts

Let us start with Hempel. The principal features of Hempel's account of explanation are (i) that explanations are arguments, (ii) that the conclusion of an expla-

I owe a long-standing debt to Peter Hempel, who first inspired my interest in the study of scientific explanation and whose writings on the topic seem to me paradigms of what is best in twentieth-century philosophy. My own thinking about explanation was redirected by Michael Friedman's seminal essay on explanation and scientific understanding, and I have also learned much from the comments, encouragement, and advice of Paul Churchland, Paul Humphreys, David Papineau, Kenneth Schaffner, and Stephen Stich. Above all I am deeply grateful to Wesley Salmon, for the depth and lucidity of his ideas and the kindness and patience of his conversation. The present essay continues a long dialogue, and, because that dialogue has been so pleasant and so instructive, I trust that it is not yet over.

nation is a sentence describing the phenomenon to be explained, and (iii) that among the premises of an explanation there must be at least one law of nature. Although the original treatment (1948) focused on cases in which the argument is deductive and the conclusion a singular sentence (a sentence in which no quantifiers occur), it was clear from the beginning that the account could be developed along two different dimensions. Thus there can be covering-law explanations in which the argument is nondeductive or in which the conclusion is general. *D-N* explanations are those explanations in which the argument is deductive and the conclusion is either a singular sentence or a nonstatistical generalization. Hempel assigned deductive explanations whose conclusion is a statistical generalization a special category – *D-S* explanations – but their kinship with the official cases of D-N explanation suggests that we should broaden the D-N category to include them (see Salmon, 1984 and this volume). Finally, *I-S* explanations are those explanations in which the argument is inductive and the conclusion a singular sentence to which the premises assign high probability.

The motivation for approaching explanation in this way stems from the character of the explanations given in scientific works, particularly in those texts that are intended to introduce students to the main ideas of various fields. Expository work in physics, chemistry, and genetics (and, to a less obvious extent, in other branches of science) often proceeds by deriving descriptions of particular events – or, more usually, descriptions of empirical regularities – from sets of premises in which statements identified as laws figure prominently. Among the paradigms, we may include: the demonstration that projectiles obtain maximum range on a flat plain when the angle of projection is 45°, the Newtonian derivation of Galileo's law of free fall, Bohr's argument to show that the frequencies of the lines in the hydrogen spectrum satisfy the formulas previously obtained by Balmer and others, the kinetic-theoretic deduction of the Boyle-Charles law, computations that reveal the energy required for particular chemical reactions, and the derivation of expected distributions of traits among peas from specifications of the crosses and Mendel's laws. In all these cases, we can find scientific texts that contain arguments that come very close indeed to the ideal form of explanation that Hempel describes.

1.2 Hempel's Problems

There are four major types of objection to the Hempelian approach. The first is the obverse of the motivational point just canvassed. Although we can identify some instances in which full-dress covering-law explanations are developed, there seem to be many occasions on which we accept certain statements as explanatory without any ability to transform them into a cogent derivation of a sentence describing the phenomenon to be explained. This objection, made forcefully in a sequence of papers by Michael Scriven (1959, 1962, 1963), includes

several different kinds of case, of which two are especially important for our purposes here. One source of trouble lies in our propensity to accept certain kinds of historical narrative – both in the major branches of human history and in evolutionary studies – as explaining why certain phenomena obtain, even though we are unable to construct any argument that subsumes the phenomena under general laws. Another results from the existence of examples in which we explain events that are very unlikely. Here the paradigm is Scriven's case (later elaborated by van Fraassen) of the mayor who contracts paresis. Allegedly, we hold that the question "Why did the mayor get paresis?" can be answered by pointing out that he had previously had untreated syphilis, despite the fact that the frequency of paresis among untreated syphilitics is low.

A second line of objection to the covering-law models is based on the difficulty in providing a satisfactory analysis of the notion of a scientific law. Hempel is especially forthright in acknowledging the problem (1965, 338). The challenge is to distinguish laws from mere accidental generalizations, not only by showing how to characterize the notion of a projectible predicate (and thus answer the questions raised by Goodman's seminal 1956) but also by diagnosing the feature that renders pathological some statements containing only predicates that are intuitively projectible (for example, "No emerald has a mass greater than 1000 kg.").

The first objection questions the necessity of Hempel's conditions on explanation. The third is concerned with their sufficiency. As Sylvain Bromberger made plain in the early 1960s (see especially his 1966), there are numerous cases in which arguments fitting one of Hempel's preferred forms fail to explain their conclusions. One example will suffice for the present. We can explain the length of the shadow cast by a high object (a flagpole or a building, say) by deriving a statement identifying the length of the shadow from premises that include the height of the object, the elevation of the sun, and the laws of the propagation of light. That derivation fits Hempel's D-N model and appears to explain its conclusion. But, equally, we can derive the height of the object from the length of the shadow, the elevation of the sun and the laws of the propagation of light, and the latter derivation intuitively *fails* to explain its conclusion. Bromberger's challenge is to account for the asymmetry.

A close cousin of the asymmetry problem is the difficulty of debarring Hempelian arguments that appeal to irrelevant factors. If a magician casts a spell over a sample of table salt, thereby "hexing" it, we can derive the statement that the salt dissolved on being placed in water from premises that include the (apparently lawlike) assertion that all hexed salt dissolves on being placed in water. (The example is from Wesley Salmon's seminal 1970; it originally comes from Henry Kyburg [1965]). But, it is suggested, the derivation does not explain why the salt dissolved.

Finally, Hempel's account of statistical explanation was also subject to special

problems. One trouble, already glimpsed in the paresis example, concerns the requirement of high probability. Among the guiding ideas of Hempel's account of explanation is the proposal that explanation works by showing that the phenomenon to be explained was to be expected. In the context of the statistical explanation of individual events, it was natural to formulate the idea by demanding that explanatory arguments confer high probability on their conclusions. But, as was urged by both Richard Jeffrey (1969) and Wesley Salmon (1970), this entails a whole class of counterintuitive consequences, generated by apparently good explanations of improbable occurrences. Moreover, the high-probability requirement itself turns out to be extremely hard to formulate (see Hempel 1965 for the surmounting of preliminary difficulties, and Coffa 1974 for documentation of residual troubles). Indeed, critics of Hempel's I-S model have charged that the high-probability requirement can only be sustained by supposing that all explanation is fundamentally deductive (Coffa 1974, Salmon 1984, 52–53).

Even a whirlwind tour of that region of the philosophical landscape occupied by theories of explanation (a region thick with syphilitic mayors, flagpoles, barometers, and magicians) can help to fix our ideas about the problems that an adequate account of scientific explanation must overcome. Contemporary approaches to the subject rightly begin by emphasizing the virtues of Hempel's work, its clarity, its connection with parts of scientific practice, its attention to the subtleties of a broad range of cases. When we have assembled the familiar difficulties, it is appropriate to ask "What went wrong?" The main extant rivals can be viewed as searching for the missing ingredient in the Hempelian approach, that crucial factor whose absence allowed the well-known troubles that I have rehearsed. I shall try to use the four main problem-types to chart the relations among Hempel's successors, and to evaluate the relative merits of the main contemporary rivals.

2. The Pragmatics of Explanation

Not all of the problem-types need be viewed as equally fundamental. Perhaps there was a basic mistake in Hempel's account, a defect that gave rise directly to one kind of difficulty. Solve that difficulty, and we may discover that the remaining troubles vanish. The suggestion is tantalizing, and it has encouraged some important proposals.

One approach is to regard the first type of problem as fundamental. Hempel clearly needed an account of the pragmatics of explanation. As his own detailed responses to the difficulties raised by Scriven (Hempel 1965, 359–64, 427) make entirely clear, he hoped to accommodate the plausible suggestion that narratives can serve an explanatory function even when we have no idea as to how to develop the narrative into an argument that would accord with one of the models. The strategy is to distinguish between what is said on an occasion in which explana-

tory information is given and the ideal underlying explanation.[1] Although the underlying explanation is to be an argument including laws among its premises, what is said need not be. Indeed, we can provide some information about the underlying argument without knowing all the details, and this accounts for the intuitions of those (like Scriven) who insist that we can sometimes say explanatory things without producing a fully approved Hempelian argument (or without knowing much about what the fully approved argument for the case at hand would be).

Instead of backing into the question of how to relate explanations to what is uttered in acts of explaining, we can take the characterization of explanatory acts as our fundamental problem. This strategy has been pursued in different ways by Peter Achinstein and Bas van Fraassen, both of whom believe that the main difficulties of the theory of explanation will be resolved by gaining a clear view of the pragmatics of explanation. Because van Fraassen's account introduces concepts that I take to be valuable to any theory of explanation, I shall consider his version.[2]

2.1 Van Fraassen's Pragmatics

Van Fraassen starts with the claim that explanations are answers to why-questions. He proposes that why-questions are essentially contrastive: the question "Why *P?*" is elliptical for "Why *P* rather than *P**, *P***, . . . ?" In this way he can account for the fact (first noted in Dretske 1973 and further elaborated in Garfinkel 1981) that the same form of words can pose different contrastive why-questions. When Willie Sutton told the priest that he robbed banks because that is where the money is, he was addressing one version of the question "Why do you rob banks?," although not the one that the priest intended.

With this in mind, van Fraassen identifies a why-question as an ordered triple $<P_k,X,R>$. P_k is the topic of the question, and an ordinary (elliptical) formulation of the question would be "Why P_k?" X is the contrast class, a set of propositions including the topic P_k. Finally R is the relevance relation. Why-questions arise in contexts, where a context is defined by a body of background knowledge K. The questions have presuppositions: each why-question presupposes that its topic is the only true member of the contrast class (intuitively, the question "Why P_k in contrast to the rest of X?" is inappropriate if P_k is false or if some other member of the contrast class is true), and also that there is at least one true proposition A that stands in the relation R to $<P_k,X>$. A why-question arises in a context K provided that K entails that the topic is the only true member of the contrast class and does not entail that there is no answer to the question (more exactly, that there is no true A bearing R to $<P_k,X>$.

Van Fraassen recognizes that the theory of explanation ought to tell us when we should reject questions rather than attempting to answer. His pragmatic ma-

chinery provides a convincing account. We reject the why-question Q in context K if the question does not arise in this context, and, instead of trying to answer the question, we offer corrections. If Q does arise in a context, then a direct answer to it takes the form "Because A," where A is a true proposition that bears R to $<P_k,X>$. The proposition A is the core of the direct answer.

2.2 Why Pragmatics Is Not Enough

Because he hopes to avoid the tangles surrounding traditional approaches to explanation, van Fraassen places no constraints on the relations that can serve as relevance relations in why-questions. In consequence, his account of explanation is vulnerable to trivialization. The trouble can easily be appreciated by noting that it is *prima facie* possible for any true proposition to explain any other true proposition. Let A, B both be true. Then, given van Fraassen's thesis that explanations are answers to why-questions, A will explain B in context K provided that there is a question "Why B?" that arises in K for which A is the core of a direct answer. We construct an appropriate question as follows: let $X = \{B, -B\}$, $R = \{<A,<B,X>>\}$. Provided that K entails the truth of B and does not contain any false proposition entailing the nonexistence of any truth bearing R to $<B,X>$, then the question $<B,X,R>$ arises in K, its topic is B, and its only direct answer is A.

Wesley Salmon and I have argued (Kitcher and Salmon 1987) that van Fraassen's account cannot avoid this type of trivialization. We diagnose the absence of constraints on the relevance relation as the source of the trouble. Intuitively, genuine why-questions are triples $<P_k,X,R>$ where R is a genuine relevance relation, and a large part of the task of a theory of explanation is to characterize the notion of a genuine relevance relation.

In fact, van Fraassen's account of the pragmatics of explanation can be used to articulate the Hempelian approach, and the articulation enables us to see how many problems remain to be resolved. Following Railton's development of Hempel's own embryonic pragmatics, let us suppose that acts of explanation provide information about underlying ideal explanatory texts. We say that $<P_k,X,R^*>$ is an *ideal* Hempelian why-question just in case R^* is the set of pairs $<A,<B,C>>$ such that (i) C is a finite set of propositions, one of which is B, (ii) A entails $B\&D$ where D is the conjunction of the negations of the remaining members of C, (iii) A contains at least one general law, and (iv) the general law is essential for the derivation of $B\&D$. An answer to an ideal Hempelian why-question is a D-N explanans for an explanandum of form $B\&D$ where B is the topic of the question and D the conjunction of the negations of the remaining members of the contrast class.[3] Now suppose that *actual* episodes of explanation involve why-questions that are *actually* answered by providing something that falls far short of an answer to an ideal why-question. Following Railton, we view

the actual answers as providing information about the answer to the ideal question. So, let us say that $<P_k,X,R>$ is a genuine why-question in context K just in case there is an ideal why-question with topic P_k and contrast class X that arises in context K such that there is an answer to $<P_k,X,R>$ that would provide those who accept K with information about an answer to the ideal why-question. Hempelian explanation-sketches are answers to genuine why-questions that do not answer the associated ideal questions.

Van Fraassen contends that his pragmatic approach to explanation solves the problem of asymmetry that arises for the Hempelian account. His solution consists in showing that there is a context in which the question "Why is the height of the tower *h*?" is answered by the proposition that the length of the shadow cast by the tower at a certain time of day is *s*. That proposition answers the question by providing information about the intentions of the builder of the tower. Thus it has seemed that van Fraassen does not touch the Hempelian problem of distinguishing the explanatory merits of two derivations (both of which satisfy the conditions of the D-N model), and that the claim to have solved the problem of asymmetry is incorrect (see Salmon 1984, 95, and Kitcher and Salmon 1987, for arguments to this effect).

The previous discussion enables us to say more precisely why van Fraassen bypasses the real problem of asymmetry. An ideal answer to the ideal why-question "Why was the height of the tower *h*?" (with some appropriate contrast class) would derive a description of the height of the tower from premises about the plans of the builder, the effectiveness of those plans, and the stability of the resulting structure. If indeed considerations of shadow length figured in the builder's plans, then it is no surprise to learn that there is a genuine why-question with the same topic and contrast class that can be answered by citing the length of the shadow. Given the right context, in which we know that shadow length was important to the builder in planning the tower, we can use the proposition about shadow length to gain information about the ideal answer to the ideal question. None of this touches the issue of why there is not an *ideal* answer to the ideal question that consists of the conjunction of the ascription of shadow length, the specification of the elevation of the sun, and the laws of the propagation of light. Translating Hempel's approach into van Fraassen's idiom, we have construed ideal Hempelian why-questions in terms of the relation R^*, and R^* allows answers that are intuitively unsatisfactory. Thus the problem of asymmetry re-emerges in van Fraassen's framework because (a) there is a serious problem of characterizing the genuine why-questions (and, correspondingly, the genuine relevance relations) and (b) the Hempelian constraints are just as inadequate in coping with this problem as they were in solving the asymmetry problem in its original guise. (*A fortiori*, an approach like van Fraassen's, that imposes no constraints on genuine relevance relations, inherits all the failures of the Hempelian approach with respect to sufficiency and more besides.)

I suggest that van Fraassen's illuminating discussion of why-questions is best seen not as a solution to all the problems of the theory of explanation, but as a means of tackling the problems of the first type (see section 1). Given solutions to the difficulties with law, asymmetry, irrelevance, and statistical explanation, we could embed these solutions in van Fraassen's framework, and thus handle the general topic of how to relate idealized accounts of explanation to the everyday practice of answering why-questions. This is no small contribution to a theory of explanation, but it is important to see that it cannot be the whole story.

2.3 Possible Goals for a Theory of Explanation

Van Fraassen's work also enables us to see how to concentrate the three residual problems that arise for Hempel's account into one fundamental issue. The central task of a theory of explanation must be to characterize the genuine relevance relations, and so delimit the class of genuine why-questions. To complete the task it will be necessary to tackle the problems of asymmetry and irrelevance, to understand the structure of statistical explanations, and, if we suppose that genuine relevance involves lawlike dependence, to clarify the concept of law.[4] However, the formulation of the task is ambiguous in significant respects. Should we suppose that there is a single set of genuine relevance relations that holds for all sciences and for all times? If not, if the set of genuine relevance relations is different from science to science and from epoch to epoch, should we try to find some underlying characterization that determines how the different sets are generated, or should we rest content with studying a particular science at a particular time and isolating the genuine relevance relations within this more restricted temporal and disciplinary area?[5]

It appears initially that Hempel sought a specification of the genuine relevance relations that was time-independent and independent also of the branch of science. However, in the light of our integration of Hempel's approach with van Fraassen's treatment of why-questions, I think we can achieve a more defensible view of the Hempelian task. Plainly, the set of ideal relevance relations (or of ideal why-questions) may be invariant across times and sciences, even though different actual questions become genuine in the light of changing beliefs. Thus one conception of the central problem of explanation – I shall call it the *Hempelian conception* – is the question of defining the class of genuine relevance relations that occur in the ideal why-questions of each and every science at each and every time. We can then suppose that variation in the why-questions arises partly from differing beliefs about which topics are appropriate, partly from differing views about the character of answers to underlying ideal why-questions, and partly from differing ideas about what would yield information about those answers.

An illustration is provided by the changing attitudes toward functional/teleological questions in biology. Consider the form of question "Why do *O*'s have *P*?"

Posed in a pre-Darwinian context, we might explicate the question in terms of a contrast class including propositions that ascribe different properties to the organisms and a relevance relation that relates propositions of the form "*P* enables *O*'s to do *X* and *doing X* promotes the welfare of *O*'s by bringing benefit *B*" to the ordered pair of topic and contrast class. We can understand the legitimacy of the question in terms of an associated why-question whose ideal answer presents the historical causes of *O*'s being as they are. In a context in which it is assumed that the history traces back to the planning of an omniscient, omnipotent, and benevolent creator, the relevance relation that figures in the functional/teleological why-question will appear appropriate because the proposition about the benefits brought by having *P* will yield information about the nature of the creator's intentions. Even though this defense is undermined after Darwin, the functional/teleological questions can still remain as legitimate. Once again, we assume that the ideal answer to the associated ideal why-question will trace the history of the origins of *O*'s. Now, however, we may assume that the having of *P* either originated or is maintained by natural selection.[6] The functional/teleological question again seems appropriate because the specification of benefits to the organism provides information about the selection pressures originating or maintaining the trait. However, it is easy to understand why functional/teleological questions disappeared from some areas of science in which they had previously seemed appropriate. Without the underlying idea that the features to be explained were directly designed or formed/maintained through a selection process, there is no obvious way to regard the specification of their beneficial effects as providing information about the ideal causal history. Hence we can see how functional/teleological why-questions lapsed in some areas of science, while, in others, they survived with an altered conception of the relationship to issues of causation.[7]

Because philosophical attention to the history of science has exposed numerous important shifts in methodological ideals, the Hempelian conception of the theory of explanation may seem far too ambitious and optimistic. However, one way to respond to claims about shifting standards is to argue that there are overarching principles of *global* methodology that apply to all sciences at all times. As particular scientific fields evolve, the principles of global methodology are filled out in different ways, so that there are genuine modifications of *local* methodology.[8] The version of the Hempelian conception that I have just sketched assigns to global methodology a characterization of ideal why-questions. Shifts in admissible why-questions, corresponding to changes in local methodology, can occur against the background of constancy in the underlying ideals – witness my brief discussion of functional/teleological questions.

Perhaps this picture makes the Hempelian conception somewhat less at odds with current thinking about the modification of methodology in the history of science. But can anything positive be said in favor of that conception? I believe

it can. The search for understanding is, on many accounts of science, a fundamental goal of the enterprise. That quest may take different forms in different historical and disciplinary contexts, but it is tempting to think that there is something that underlies the various local endeavors, something that makes each of them properly be seen as a striving after the same goal. The Hempelian conception proposes that there is an abstract conception of human understanding, that it is important to the development of science, and that it is common to the variety of ways in which understanding is sought and gained. Scientific explanations are intended to provide objective understanding of nature. The task of characterizing the ideal notions of explanation, why-question, and relevance is thus one of bringing into focus one of the basic aims of science.

I do not suppose that these remarks provide any strong reasons for thinking that the Hempelian conception is correct. It might turn out that there is nothing but ritual lip movements in the avowal of explanation as an aim of the sciences. Nonetheless, there is an obvious motivation for pursuing the Hempelian conception, for, if it is correct, then we can hope to obtain some insight into the rationality and progressiveness of science. Since I know of no conclusive reasons for abandoning my preferred version of the conception, I propose to consider theories of explanation that undertake the ambitious task of characterizing the ideal relevance relations. More modest projects can come once ambition has failed.

3. Explanation as Delineation of Causes

There are two main approaches to explanation that can be seen as undertaking the project just outlined. One of these can be motivated by considering the problems of asymmetry. Intuitively, the length of the shadow cast by a flagpole is causally dependent on the height of the flagpole, but the height is not causally dependent on the shadow-length. Thus we arrive at the straightforward proposal that Hempel's failure to solve problems of asymmetry (and irrelevance) stems from the fact that causal notions are avoided in his analyses. Diagnosis leads quickly to treatment: genuine relevance relations are causal relations, explanations identify causes.

Of course, the invocation of causal notions has its costs. Hempel's account of explanation was to be part of an empiricist philosophy of science, and it could therefore only draw on those concepts that are acceptable to empiricists. If causal concepts are not permissible as primitives in empiricist analyses, then either they must be given reductions to empiricist concepts or they must be avoided by empiricists. Hempel's work appears to stand in a distinguished tradition of thinking about explanation and causation, according to which causal notions are to be understood either in terms of the concept of explanation or in terms of concepts that are themselves sufficient for analyzing explanation. Empiricist concerns about the evidence that is available for certain kinds of propositions are frequently trans-

lated into claims about conceptual priority. Thus, the thesis that we can only gain evidence for causal judgments by identifying lawlike regularities generates the claim that the concept of law is prior to that of cause, with consequent dismissal of analyses that seek to ground the notion of law in that of cause.

One of Hume's legacies is that causal judgments are epistemologically problematic. For those who inherit Hume's theses about causation (either his positive or his negative views) there are obvious attractions in seeking an account of explanation that does not take any causal concept for granted. A successful analysis of explanation might be used directly to offer an analysis of causation—most simply, by proposing that one event is causally dependent on another just in case there is an explanation of the former that includes a description of the latter. Alternatively, it might be suggested that the primitive concepts employed in providing an analysis of explanation are just those that should figure in an adequate account of causation.

Because the invocation of causal dependency is so obvious a response to the problems of asymmetry and irrelevance, it is useful to make explicit the kinds of considerations that made that response appear unavailable. One central theme of the present essay is that there is a tension between two attractive options. Either we can have a straightforward resolution of asymmetry problems, at the cost of coming to terms with epistemological problems that are central to the empiricist tradition, or we can honor the constraints that arise from empiricist worries about causation and struggle to find some alternative solution to the asymmetries. The two major approaches to explanation respond to this tension in diametrically opposite ways. As we may anticipate, the central issues that arise concern the adequacy of proposed epistemological accounts of causation and of suggestions for overcoming problems of asymmetry and irrelevance without appealing to causal concepts.

Before we pursue these questions, it will help to have a more detailed view of both approaches. The remainder of this section will be devoted to the causal approach. I shall examine its rival in section 4.

3.1 Causal Why-Questions and Causal Explanations

Let us start with the explanation of particular facts and events. We restrict our attention to singular propositions that describe particular facts and events—paradigmatically such things as the mayor's having paresis or an electron's tunneling through a potential barrier. Call any why-question that has a singular proposition as its topic a *singular* why-question. An admissible contrast-class for a singular why-question is a set of propositions, among which is the topic of the why-question, such that the propositions exhibit *homogeneous variation* with respect to some property or object described in the topic. A set of propositions exhibits homogeneous variation just in case each pair of propositions in the set has some common constituent (property or object) and the common constituents of

any two propositions in the set are the same. This restriction of contrast-classes is intended to permit such classes as {"Sutton robbed the bank," "Sutton robbed the grocery store," "Sutton robbed the church"} and {"Sutton robbed the bank," "Sutton worked as a mechanic," "Sutton worked as a neurosurgeon"} but to debar such bizarre collections as {"Sutton robbed the bank," "Shakespeare died in 1916," "Babe Ruth hit more home runs than Hank Aaron"}.

An ideal singular why-question is a triple $<P,X,R>$ where P is a singular proposition, X is an admissible contrast-class (including P), and R is the relation of *complete causal relevance*. This relation obtains between A and $<P,X>$ just in case A is the proposition expressing all and only the causal information relevant to the truth of P and the falsity of the other members of X. Intuitively, A tells the complete causal story about why P is true and the other members of the contrast-class are false. Although there may be room for skeptical worries about the coherence of the notion of a complete causal story, I shall suppose that there is indeed some proposition that relates that part of the history of the universe culminating in the obtaining of the state decribed by P and in the nonobtaining of the states described by the other members of the contrast-class.[9]

Actual why-questions are rarely aimed at eliciting the entire causal history underlying a particular event or state. Those who ask why typically want to elicit a certain type of information about the causal antecedents of the event/state described in the topic. So I shall take an actual singular why-question to be a triple $<P, X, R>$ where P is a singular proposition, X an admissible contrast-class, and R a relation of *particular* causal relevance. Each particular causal relevance relation R is associated with some condition C such that A bears R to $<P,X>$ just in case (i) A is a logical consequence of the proposition A^* that bears the relation of complete causal relevance to $<P,X>$, (ii) A satisfies C, and (iii) A is not a logical truth. I shall assume that no restrictions need to be placed on the appropriate associated conditions C. In effect, this allows for genuine actual why-questions that are directed at eliciting information about any aspect of the causal history behind an event or state. Typically, we focus on some temporal period that interests us and on some state that obtained (or some event that occurred) during this period. There is an actual why-question, whose topic describes the height of the tower, that is answered by citing the length of the shadow because it is pertinent to inquire about the intentions of the builder at the time the tower was designed. More formally, we ask for a consequence of the complete causal history that describes the desired effect that a tower of the given height would have.

As I have emphasized, an obvious virtue of the causal approach is that it handles problems of asymmetry and irrelevance. How does it fare with respect to the other difficulties that beset the Hempelian account? Some of the objections raised by Scriven evaporate immediately once we have an adequate pragmatics of explanation (as indeed Hempel pointed out in his own seminal discussions of them). Others require us to make explicit a point about causal determination.

Consider the case of the mayor's paresis. If we inquire why the mayor, rather than some one of the other townspeople, contracted paresis, then an adequate answer may be to point to the mayor's prior history of untreated syphilis. Here, the topic of our question is the proposition that the mayor contracted paresis, the contrast-class contains propositions ascribing paresis to the other (more fortunate) citizens, and the relevance relation is a particular relation of causal relevance based on the associated condition of describing some antecedent feature of the mayor's medical state that distinguishes him from the other townsfolk. *It is not required that identification of that prior state should enable us to deduce that the mayor later contracted paresis, or to infer with high probability that he would later contract paresis, or even that it should raise the prior probability of his contracting paresis.* There is no demand that answers should make the topics of the question more expectable than they were previously.[10]

It is not hard to see that the causal approach bypasses the problem of providing an analysis of scientific laws. While proponents of the approach may believe that, in general, complete causal histories will mention laws of nature—or even that the structure of these histories may sometimes accord with the requirements of the covering-law models—it is not incumbent on them to provide an analysis of the notion of law, at least not for the purposes of giving an account of scientific explanation. Moreover, they may even remain agnostic with respect to the question whether all ideal explanations (ideal answers to why-questions) involve general laws.

The causal approach also has little difficulty in overcoming the problems that confronted Hempel's account of statistical explanation. As already indicated in the discussion of the paresis case, the approach has no commitment to the idea that there is any statistical relationship between the information given in an answer to a why-question and the topic (or ordered pair of topic and contrast-class) of the why-question. What is important is the provision of causal information, and once this has been achieved the effects on the probability of the topic (or on other members of the contrast class) is of no concern.[11]

These successes for the causal approach are impressive, and they provide some motivation for considering whether the traditional empiricist arguments that dissuaded Hempel (and others) from using causal concepts in the analysis of explanation are cogent. We shall pursue this question in some detail below. For the moment, however, it is necessary to examine whether or not the causal approach is vulnerable to different kinds of objection.

3.2 Are There Noncausal Explanations of Singular Propositions?

One important worry about the causal approach is that, as I have so far characterized it, it is restricted to giving an account of the explanation of singular propositions. Before we try to remedy this deficiency, it is useful to consider whether

there are some cases in which singular propositions are explained in noncausal ways. More exactly, we should ask if it is always true that the ideal answer to an ideal question with a singular topic is the associated complete causal history. The doubt divides into two parts: (i) are there singular propositions that give rise to why-questions and that can be explained by answering those why-questions, but which describe phenomena that do not have causal histories?; (ii) are there singular propositions that describe phenomena which have causal histories and which give rise to why-questions, but which are ideally explained without relating the complete causal histories?

There are areas of inquiry, formal sciences, in which investigators appear to put forward explanations despite the fact that the phenomena to be explained do not have causes. Two obvious examples are formal syntax and pure mathematics. Explanations of the grammaticality or ungrammaticality of particular strings in particular natural languages are given by identifying the constraints set by the underlying rules of syntax.[12] Note that it will not do to suggest that the formal explanation is a placeholder for a description of causal processes that occur in the brains of speakers – for part of the point of the enterprise is to distinguish between explanations of competence and explanations of performance.

I shall not pursue the examples from formal linguistics, because it seems to me that, in the case of mathematics there are several instances in which we can distinguish between the explanatory worth of arguments that yield a particular theorem, even though we cannot claim that one of the arguments provides insight into the *causes* of the fact reported in the theorem. Hence, if this is indeed correct, the problems that provoke the introduction of causal concepts in the theory of explanation for singular propositions about physical states of affairs (paradigmatically problems of asymmetry and irrelevance) have counterparts in domains to which causal notions are inapplicable.

Before examining some examples, I shall try to forestall an objection, forcefully presented to me by Paul Humphreys. Can we legitimately talk about explanation in mathematics, and, even if we can, should we suppose that attention to mathematical cases will indicate anything about *scientific* explanation? I reply to the skepticism that provokes the question in two ways. First, given my own views of the nature of mathematics (see Kitcher 1983, 1987b, 1988) mathematical knowledge is similar to other parts of scientific knowledge, and there is no basis for a methodological division between mathematics and the natural sciences. Second, the importance of mathematical explanation in the growth of mathematical knowledge is appreciated not only by philosophers who do not share my heterodoxies (see, for example, Steiner 1978) but also by mathematicians when they are engaged in critical discussion of specific mathematical reasoning. The examples that follow are intended to show that proofs and axiomatizations are, at least sometimes, assessed for their explanatory merits, and it seems to me that the bur-

den of argumentation is on skeptics who wish to campaign for some sort of methodological dualism.

A. The theorem that a function with derivatives of all orders that takes values of opposite sign at the endpoints of an interval has a zero within the interval (the *intermediate zero theorem*) can be justified as follows. Any function meeting the conditions given can be represented as a smooth curve, in the intuitive sense of a curve that can be drawn without lifting the pencil from the paper. Since the curve must lie below the axis at one end of the interval and above the axis at the other end of the interval, there must be a point within the interval at which it crosses the axis. This point corresponds to a zero of the function. Bolzano complained that this argument inverts the true order of the sciences (see Kitcher 1975 for exegesis), and he set out to find a more satisfactory derivation. While Bolzano recognized that the appeal to the properties of smooth curves produces conviction *that* the theorem is true, he contended that it fails to identify "the reason for the fact" (in Aristotle's famous terminology), and he suggested that the failure stems from the use of considerations that are extraneous to claims about numbers and functions. Instead, he sought, and Weierstrass later completed, a proof that would identify the properties of the real line on which the intermediate zero theorem depends. I suggest that the problem posed by Bolzano is a mathematical analogue of the irrelevance problem for explanations in the natural sciences, and that we cannot diagnose the flaw in the intuitive geometrical argument by suggesting that geometrical considerations are *causally* irrelevant to truths about numbers and functions. Indeed, Bolzano's insight is that there is a broader notion of objective dependency to which correct explanations must conform—an insight that derives ultimately from Aristotle.

B. It is possible to axiomatize the theory of finite groups in a number of different ways. The standard approach takes a group to be a finite set which is closed under an associative operation, *multiplication*. There is an idempotent element, 1, such that, for any element a of the group $a1 = 1a = a$; and, for any element a there is an inverse, a^{-1} such that $a^{-1}a = aa^{-1} = 1$. On this basis one can prove that division is unique wherever it is possible. Alternatively, a finite group can be identified as a finite set, closed under an associative operation, *multiplication*, such that division is unique wherever it is possible. On this basis, one can show that finite groups have all the properties attributed in the usual axioms. Mathematicians distinguish the two axiomatizations—if only in their practice of choosing the standard axioms—and they sometimes express the preference by saying that the usual axioms are more "natural" than the nonstandard ones, or that the division property is "less fundamental." I suggest that what they are recognizing is a case of the asymmetry problem: we can explain why finite groups satisfy the division property by using the axioms about the existence of inverses and idempotent elements to demonstrate that division is unique wherever it is possible. But the derivation of the existence of an idempotent element and of inverses

from the division property is nonexplanatory, and, I think, nonexplanatory in just the same way as the derivation of heights from shadowlengths. If this is correct, then the example reveals that the asymmetry problem can arise in cases where causal considerations are quite beside the point. Moreover, in this case and in that discussed in A, it is not hard to see a reason for the distinguishing of the derivations: the preferred derivation can be generalized to achieve more wide-ranging results. Thus, by using the ideas developed by Bolzano, Weierstrass, and Dedekind, we can show that the intermediate zero theorem holds in the much weaker case in which the function is merely continuous. Similarly, if we drop the restriction to *finite* groups, we can show that any group satisfying the standard axioms has the division property, but it is not the case that any set closed under an associative operation with the division property has an idempotent element and inverses (that claim need not hold when the set is infinite). In both instances, the explanatory derivation is similar to derivations we could provide for a more general result; the nonexplanatory derivation cannot be generalized, it applies only to the local case. I shall try to show later how this diagnosis can be made more precise, and how we can use it to offer a picture of explanation that differs from the causal approach.

C. There are numerous classes of equations in one variable for which we can specify solutions as rational functions of the coefficients. This is trivial in the case of the class of linear equations ($ax + b = 0$), and familiar in the case of quadratic equations ($ax^2 + bx + c = 0$). It also holds for cubic equations and for quartic equations. To solve the general cubic, $ax^3 + bx^2 + cx + d = 0$, it suffices to note that setting $x = z - b/3a$, generates an equation of the form

$$(*)\ pz^3 + qz + r = 0$$

where p, q, and r are all rational functions of a, b, and c. If we now set $z = y - q/3py$, and substitute in (*), we obtain an equation of the sixth degree whose coefficients are rational functions of p, q, r, (and therefore of a, b, and c), and which is a quadratic in y^3. Here we can apply the ordinary formula for quadratic equations to obtain an expression for y, and the expression of z, and x, follows in two easy steps.

Since a similar trick works for the quartic equation, we can show, for each class of polynomial equations up to and including degree 4 that the roots can be written as rational functions of the coefficients. So much was appreciated by the end of the eighteenth century (in fact, even earlier), but mathematicians concerned with the theory of equations—most notably Lagrange—believed that the mere ability to provide the derivations alluded to in the last paragraph does not show us why equations in these classes permit expression of the roots as rational functions of the coefficients. Some insight into the structure behind the phenomenon is needed, and this was provided partially by Lagrange's investigations of the effects of permutations of the roots on various functions in the roots, ultimately

in Galois's development of the theory that bears his name. After Galois, we have a criterion for the expressibility of roots as rational functions of the coefficients, to wit the solubility of the Galois group of the equation, and we can see just why this applies in the four special cases. Nonexplanatory special derivations give way to an explanatory proof drawn from a general theory about the properties of classes of equations.

I turn now to the second type of concern about the causal approach to the explanation of singular propositions. As we shall see, this worry has affinities with the complaint made in C above, namely that why-questions are frequently posed in a search for theoretical understanding. In the case described in C, there was no issue of finding causal histories of the phenomena to be explained. However, even when causal histories are available, they may not be what the explanation requires. Two examples will illustrate the moral.

D. There is a party trick in which someone "knots" a telephone cord around a pair of scissors. In fact, no genuine knot is produced, and the scissors can easily be removed (and the cord returned to its standard configuration) if the victim makes a somewhat unobvious twist at the start. Those who do not make the right initial twist can struggle for hours without getting anywhere. What explains their failure? In any such case, we could, of course, provide the causal details, showing how the actions actually performed lead to ever more tangled configurations. But this appears to omit what should be central to the explanation, namely the fact that the topological features of the situation allow for disentanglings that satisfy a specifiable condition, so that sequences of actions which do not satisfy that condition are doomed to failure. We need to know the topological structure that lies behind the vicissitudes of the particular attempt and the particular failure.[13]

E. We discover that, for a particular city, over a period of a century, the sex-ratio at birth, combined over all the hospitals, is always very close to 1.04 to 1, with males being more common. There is a complete causal history underlying this fact: it involves vast numbers of details about the production of sperm and eggs, circumstances of mating, intra-uterine events, and so forth. However, in explaining the sex-ratio, we do not want any of this information. Instead, it suffices to point out that there are selection pressures on individuals of *Homo sapiens* that result in the approximate attainment of a 1–1 sex ratio at reproductive age, and that higher male mortality between birth and reproduction requires a natal sex-ratio of 1.04 to 1. Having a 1:1 sex-ratio at reproductive age is an evolutionary equilibrium for a species like ours, and we explain demographic data from a large local population by showing how they approximate the evolutionary equilibrium.[14]

In both instances, the causal approach seems to err by overlooking the fact that the particular phenomenon to be explained is one example of a class, all of whose members instantiate a general regularity. Post-Hempelian philosophers of science have sometimes delighted in attacking the covering-law models by noting

that the mere citation of a covering law is often a very poor explanation: if we hope to explain why a particular sample of copper expanded when heated, then we gain very little from the statement that all samples of copper expand when heated. But, in examples D and E, the identification of the regularity is an ingredient in the explanation. We explain a victim's frustration with the telephone cord by identifying the topological features of the "knot," and noting that only certain kinds of actions will produce the desired result. We explain the birth sex-ratio by offering a general claim about sex-ratios in large populations of *Homo sapiens*, and then going on to explain why that claim holds. To accommodate both the features of these examples and the legitimacy of the complaint about the triviality of providing covering laws, I suggest that singular why-questions are often concerned to relate the phenomenon described by the topic to other similar phenomena, rather than to fathom the causal details of a particular situation. In some cases, the delineation of a class of instances in which similar things happen is only the first stage in explanation, for the original question was concerned with why the phenomenon is to be found throughout the class. In such cases, the intent of the question is implicitly general and we could say that, while the *apparent* topic of the question is singular, the *real* topic concerns a regularity. Hence, merely stating the regularity is explanatorily worthless and the poverty of the simplest forms of covering-law derivation can easily be understood. However, there are other occasions on which the identification of a phenomenon as belonging to a class—typically defined in terms of some language that does not occur in the initial posing of the question—suffices to explain it. The failure to untangle the telephone cord is explained by using topological notions to characterize the initial configuration of the wire. Or the delineation of the class can be a preliminary step in explanation, as when we formulate the general regularity about population sex-ratios in *Homo sapiens* and then proceed to the evolutionary explanation.

The negative point of these examples is that the account of singular explanation needs to be amended to allow that explanations need not, and sometimes should not, deliver information about the causal history of a particular occurrence. The positive point concerns the penetration of singular explanation by theoretical explanation. It might be tempting to believe that the explanation of singular propositions can be studied autonomously without worrying about the character of theoretical explanation. Examples D and E are intended to show that this is an illusion. Even when we are interested in explaining a particular event or state, the explanation we desire may well be one that would also explain something quite general, and any attention to the local details may be misguided and explanatorily inadequate. The same point is made in the mathematical context by example C.

Proponents of the causal approach can respond by revising their account of genuine relevance relations and genuine why-questions. For example, the conditions on genuine relevance relations can now be viewed as disjunctive allowing either for answers that provide information about the complete causal history or

for answers that provide information about constraints on causal processes of a particular kind, to wit processes that generate phenomena similar (in some specified way) to the phenomenon reported in the topic. The task of working this out in full detail is not trivial, but, even were it to be done successfully, there would still be concern about the liberality of the proposal. Although the explanations I have claimed to be preferable in cases like D and E would now be accommodated, the recitations of causal detail would not have been debarred. Whether they should be seems to me to be an interesting open question.

However, the issue of how to integrate an account of causal singular explanation and an account of theoretical explanation (the issue bruited in the last paragraph) appears less fundamental than the problem of providing a characterization of theoretical explanation itself. The plausibility of the causal approach derives chiefly from its handling of singular propositions – as we turn our attention to the explanation of general propositions, the talk of causation comes to seem forced or even inapplicable.

3.3 Causal Explanation and Theoretical Explanation

Theoretical explanation provides some support for the Hempelian idea that explanation is derivation. For, when we consider the paradigms of theoretical explanation, the Newtonian derivation of Kepler's laws, quantum chemical accounts of the propensities of elements for forming compounds, molecular biology's explanation of the copying of genetic material, plate tectonic accounts of the presence of earthquake zones, the derivations found in standard sources seem to provide ideal explanations. Nonetheless, Hempel's account of theoretical explanation is underdeveloped, and with good reason. As Hempel and Oppenheim clearly recognized (1948, note 33), if theoretical explanation is conceived in terms of the derivation of laws using laws as premises, then it is unclear why we cannot explain any law by deriving it from the conjunction of itself with any other law. The obvious response to the difficulty is to say that we explain laws by deriving them from more fundamental laws. If the causal approach to explanation is to be fully developed, it must provide some way of saying what is meant by the intuitive (but murky) thesis that some laws are more fundamental than others.

Consider the quantum mechanical account of the characteristics of the periodic table. At the first (and most informal) step one introduces the idea of discrete energy levels in the atom and applies the Pauli exclusion principle (see, for a classic source, Pauling 1960, 47ff.). This provides an attribution to each element of the pattern of shell-filling around an atomic nucleus, and the sequence reveals the periodicity originally established by Mendeleev. Thus, to cite one example, we discover that the noble gases (helium, neon, argon, and so forth) have in common the property that their outermost electron shells (i.e., the highest energy levels in which electrons occur around the nucleus) are completely filled. Their stability

is then interpreted in terms of the lack of opportunity for the formation of ionic bonds (involving transfer of electrons) or covalent bonds (sharing of electrons). The characterizations thus achieved are subsequently obtained by applying the formalism of quantum mechanics to provide a rigorous account of shell-filling, of stability, and of bond-formation.

The generalization that noble gases do not form compounds with other elements is explained by beginning with the laws governing bond formation, deriving the conclusion that bond formation requires complementation with respect to the filling of the outermost electron shells, concluding, as a special case, that when the outermost shells are already filled there are no opportunities for complementation in terms of electron transfer or electron sharing, and finally using the law that noble gases have their outermost shells filled to infer that noble gases do not form compounds with other elements. The derivation can be enriched by treating one (or more) of the generalizations employed in it from a formal quantum-mechanical perspective (this amounts to what I shall call *explanation extension*—see section 4.5 below). How is the explanatory power of the derivation to be understood from the perspective of the causal approach?

For those, like Salmon, who think of the explanation of singular propositions as primary, the obvious response is to suggest that the explanation of regularities involves the identification of mechanisms that are at work in all the cases covered by the regularity. The problem is to make this formulation more precise. What is it for a "mechanism to be at work in an event, state, or process"? What is it for an event, state, or process to be "covered by the regularity"? Let us look at the difficulties involved in clarifying these notions by considering in more detail the example discussed in the last paragraph.

The states and events covered by the law that noble gases are chemically inert are, it would appear, episodes in which samples of noble gas occur in the presence of samples of other elements (or of compounds). On each occasion we can recognize a time at which the noble gas molecules first enter the scene, and we can imagine a complete causal history that traces the interrelations between these molecules and molecules of the other substances that are present. Now we ask what makes the shell-filling properties of the noble gas molecules the fundamental mechanism that accounts for the fact that, in all these episodes, we find no chemical combination between the noble gas sample and the other substances. Those properties are, to be sure, one aspect of all the causal histories—as is the fact that no bonds form—but it is hard to see what distinguishes it as crucial to the explanation.

The obvious response is to claim that we have a general account of what occurs when elements (or compounds) *do* combine, an account that makes reference to electron transfer, electron sharing, and shell-filling. This general account reveals to us that the shell-filling properties of the noble gases operate as a constraint on the causal processes that can occur when they are brought into the presence of

other substances. What counts as a fundamental mechanism is not dictated by the local details of the individual causal processes that occur in these episodes. Rather, the fundamental mechanisms are disclosed by our most general theoretical accounts of a range of regularities. The task of understanding the explanatory adequacy of these general accounts thus remains as a presupposition of the causal approach, and I do not see that it can be completed in terms of the local studies of individual cases of causation that the causal approach takes as primary.

In an earlier essay (Kitcher 1985c), I contrasted "top down" and "bottom up" approaches to explanation. That contrast resurfaces in the present context. Top down approaches will attempt to provide an account of what theoretical explanation is, use this as a basis for underwriting talk about "fundamental mechanisms," and so proceed toward the identification of causes in particular cases. Bottom up approaches view us as having the ability to discern causal relations in specific episodes, and see theoretical explanation as stitching together results about the causation of individual states and events. The considerations I have just advanced are attempts to show that the project of proceeding from singular explanation to theoretical explanation is more problematic than one might have thought. Those considerations obtain greater force in light of the earlier point that theoretical explanation penetrates singular explanation. Finally, waiting in the wings is the complaint that explanation can go forward in areas of discourse in which causal notions are inapplicable and that these areas sometimes order the phenomena in ways that have similar features to the explanatory orderings of the natural sciences.

Nonetheless, as I emphasized earlier in this section, the causal approach has obvious merits. Does it have any serious rival? Let us see.

4. Explanation as Unification

On both the Hempelian and the causal approaches to explanation, the explanatory worth of candidates—whether derivations, narratives, or whatever—can be assessed individually. By contrast, the heart of the view that I shall develop in this section (and which I shall ultimately try to defend) is that successful explanations earn that title because they belong to a set of explanations, the *explanatory store*, and that the fundamental task of a theory of explanation is to specify the conditions on the explanatory store. Intuitively, the explanatory store associated with science at a particular time contains those derivations which collectively provide the best systematization of our beliefs. Science supplies us with explanations whose worth cannot be appreciated by considering them one-by-one but only by seeing how they form part of a systematic picture of the order of nature.

4.1 The Ideal of Unification

All this is abstract and somewhat metaphorical. To make it more precise, let us begin with the proposal that *ideal* explanations are derivations. Here there is

both agreement and disagreement with Hempel. An argument can be thought of as an ordered pair whose first member is a set of statements (the premises) and whose second member is a single statement (the conclusion). Hempel's proposal that explanations are arguments appears to embody this conception of arguments as *premise-conclusion* pairs. But, on the systematization account, an argument is considered as a derivation, as a sequence of statements whose status (as a premise or as following from previous members in accordance with some specified rule) is clearly specified. An ideal explanation does not simply list the premises but shows how the premises yield the conclusion.

However, the systematization approach retains the Hempelian idea that to explain a phenomenon is to produce an argument whose conclusion describes the phenomenon, and this would appear to founder on the difficulties adduced by Jeffrey (1969) and Salmon (1984) concerning the explanation of objectively improbable events. I shall postpone discussion of this point to the next section.

For a derivation to count as an *acceptable* ideal explanation of its conclusion in a context where the set of statements endorsed by the scientific community is K, that derivation must belong to the explanatory store over K,$E(K)$. At present, I shall assume that K is both consistent and deductively closed, and that the explanatory store over a set of beliefs is unique. $E(K)$ is to be the set of derivations that best systematizes K, and I shall suppose that the criterion for systematization is unification.[15] $E(K)$, then, is the set of derivations that best unifies K. The challenge is to say as precisely as possible what this means.

We should be clear about just what is to be defined. The set of derivations we are to characterize is the set of explanations that would be acceptable to those whose beliefs comprised the members of K. At this stage, the project does not provide an account of *correct* explanation, and it will be important to remedy that deficiency later. To avoid metaphysical complications, my attempt will be postponed to the final section.

The idea that explanation is connected with unification has had some important advocates in the history of the philosophy of science. It appears to underlie Kant's claims about scientific method[16] and it surfaces in classic works in the logical empiricist tradition (see Hempel [1965] 345, 444; Feigl [1970] 12). Michael Friedman (1974) has provided the most important defense of the connection between explanation and unification. Friedman argues that a theory of explanation should show how explanation yields understanding, and he suggests that we achieve understanding of the world by reducing the number of facts we have to take as brute.[17] Friedman's motivational argument suggests a way of working out the notion of unification: characterize $E(K)$ as the set of arguments that achieves the best tradeoff between minimizing the number of premises used and maximizing the number of conclusions obtained.

Something like this is, I think, correct. Friedman's own approach did not set

up the problem in quite this way, and it proved vulnerable to technical difficulties (see Kitcher 1976 and Salmon, this volume). I propose to amend the account of unification by starting from a slight modification of the motivational idea that Friedman shares with T. H. Huxley (see note 17). Understanding the phenomena is not simply a matter of reducing the "fundamental incomprehensibilities" but of seeing connections, common patterns, in what initially appeared to be different situations. Here the switch in conception from premise-conclusion pairs to derivations proves vital. *Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and, in demonstrating this, it teaches us how to reduce the number of types of facts we have to accept as ultimate (or brute).*[18] So the criterion of unification I shall try to articulate will be based on the idea that *E(K)* is a set of derivations that makes the best tradeoff between minimizing the number of patterns of derivation employed and maximizing the number of conclusions generated.

4.2 Argument Patterns

First we need the notion of pattern. A *schematic sentence* is an expression obtained by replacing some, but not necessarily all, the nonlogical expressions occurring in a sentence with dummy letters. Thus, starting with the sentence "Organisms homozygous for the sickling allele develop sickle-cell anemia," we can generate a number of schematic sentences: for example, "Organisms homozygous for *A* develop *P*" and "For all *x*, if *x* is *O* and *A* then *x* is *P*" (the last being the kind of pattern of interest to logicians, in which *all* the nonlogical vocabulary gives way to dummy letters). A set of *filling instructions* for a schematic sentence is a set of directions for replacing the dummy letters of the schematic sentence, such that, for each dummy letter, there is a direction that tells us how it should be replaced. For the schematic sentence "Organisms homozygous for *A* develop *P*," the filling instructions might specify that *A* be replaced by the name of an allele and *P* by the name of a phenotypic trait. A *schematic argument* is a sequence of schematic sentences. A *classification* for a schematic argument is a set of statements describing the inferential characteristics of the schematic argument: it tells us which terms of the sequence are to be regarded as premises, which are inferred from which, what rules of inference are used, and so forth. Finally, a *general argument pattern* is a triple consisting of a schematic argument, a set of sets of filling instructions, one for each term of the schematic argument, and a classification for the schematic argument.

A particular derivation, the sequence of sentences and formulas found in a scientific work for example, instantiates a general argument pattern just in case: (i) the derivation has the same number of terms as the schematic argument of the

general argument pattern, (ii) each sentence or formula in the derivation can be obtained from the corresponding schematic sentence in accordance with the filling instructions for that schematic sentence, (iii) the terms of the derivation have the properties assigned by the classification to corresponding members of the schematic argument. Later in this section I shall offer some examples that are intended to show how this conception of argument patterns functions in actual cases to capture the structure that underlies the derivations put forward within particular scientific fields.

Derivations may be similar either in terms of their logical structure or in terms of the nonlogical vocabulary they employ at corresponding places. The notion of a general argument pattern allows us to express the idea that derivations similar in either of these ways have a common pattern. However, similarity is a matter of degree. At one extreme, a derivation is maximally similar to itself and to itself alone; at the other, any pair of arguments can be viewed as having a common pattern. To capture the notion that one pair of arguments is more similar than another pair, we need to recognize the fact that general argument patterns can demand more or less of their instantiations. If a pattern sets conditions on instantiations that are more difficult to satisfy than those set by another pattern, then I shall say that the former pattern is more *stringent* than the latter.

The stringency of an argument pattern is determined in part by the classification, which identifies a logical structure that instantiations must exhibit, and in part by the nature of the schematic sentences and the filling instructions, which jointly demand that instantiations should have common nonlogical vocabulary at certain places. If both requirements are relaxed completely then the notion of pattern degenerates so as to admit of *any* argument. If both conditions are simultaneously made as strict as possible then we obtain another degenerate case, a "pattern" which is its own unique instantiation. Relaxing the demands on nonlogical vocabulary (the conditions set by the schematic sentences and the filling instructions) while also requiring that the classification determine the precise inferential status of each term in the schematic argument yields the logician's notion of pattern.

Plainly, we make the global problem more tractable if we compare the relative stringency of argument patterns that differ only with respect to their classifications and assess relative stringency where the classifications are the same and the differences are confined to the demands made by the schematic sentences and the filling instructions. The general problem would require us to make comparisons in cases where the classification of P_1 sets more exacting conditions than the classification of P_2 but the asymmetry is reversed with respect to the conditions on substituting dummy letters. In this essay, I shall not try to offer a general account of relative stringency. As we shall discover in section 7, the relevant comparisons often involve only the straightforward cases.

4.3 Systematization of Belief

We want *E(K)* to be the set of arguments that best unifies *K*. Typically, there are many ways of deriving some statements of *K* from others. Call any set of arguments that derives some members of *K* from other members of *K* a *systematization* of *K*. *E(K)* will be the best systematization of *K*.

The initial requirement that must be met by *E(K)* is that all the arguments it contains should be acceptable relative to *K*. Say that a set of derivations is acceptable relative to *K* just in case each step in each derivation is deductively valid and each premise of each derivation belongs to *K*. In considering ways of systematizing *K* we restrict our attention to those sets of derivations that are acceptable relative to *K*. This is an idealization, since we often admit as explanatory arguments that do not accord with our present beliefs: for example, in using Newtonian dynamics to explain trajectories or Mendelian rules to account for a distribution of traits. I shall show how the idealization can be relaxed in the next section.

Continuing with the central ideas of the account, let us define a *generating set* for a set of derivations to be a set of argument patterns such that each derivation in the set instantiates some pattern in the generating set. If we have a set of derivations *D* and a generating set for it *G*, then *G* will be said to be complete with respect to *K* just in case every derivation that is acceptable relative to *K* and which instantiates a pattern in *G* belongs to *D*. In determining the explanatory store *E(K)* we first narrow our choice to those sets of arguments that are acceptable relative to *K*, the acceptable systematizations of *K*. Then we consider, for each such set of derivations, the various generating sets that are complete with respect to *K*. (The importance of the requirement of completeness is to debar explanatory deviants who use patterns selectively. If someone claims that an argument instantiating a particular pattern explains why Mars follows the trajectory it does, admits that there is an acceptable derivation instantiating the same pattern that will yield as its conclusion a description of the trajectory of Venus, but refuses to allow the latter derivation as explanatory, then, I suggest, that person has incoherent views about explanation.) Now, having associated with each acceptable systematization a collection of complete generating sets that generates it, we pick out for each acceptable systematization that member of the associated collection of complete generating sets that does best according to criteria for unification (to be indicated shortly). Call this the *basis* of the systematization in question. Finally, we rank the bases of the acceptable systematizations in terms of their unifying power. *E(K)* is that acceptable systematization whose basis ranks highest.

The intuitive idea behind unification is the generation of as many conclusions as possible using as few patterns. It is also important that the instantiations of the patterns should genuinely be similar, that is, that the patterns in question should be stringent. With this in mind, define the *conclusion* set of a set of derivations *D*, *C(D)*, to be the set of statements that occur as conclusions of some member

of D. The unifying power of a complete generating set for D varies directly with the size of $C(D)$, directly with the stringency of the patterns in the set, and inversely with the number of patterns in the set. As in the case of our discussion of stringency, I shall not explore the ways in which tradeoffs among these factors might be made. I am prepared to allow for the possibility that, with respect to some possible corpora K, there might be genuine indeterminacy in deciding how to weigh relative stringency, paucity of patterns and range of conclusions against one another, with consequent indeterminacy about $E(K)$. But I shall argue below that with respect to actual (present or past) sets of beliefs it is possible to use the incomplete criteria I have given to judge the merits of rival systematizations and so resolve the difficulties that problems of asymmetry and irrelevance pose for theories of explanation.

4.4 Why-Questions Revisited

The account that I have offered needs to be integrated with the approach to pragmatic issues given in section 2.1. We can proceed in a fashion parallel to that adopted in articulating the causal approach (3.1), with the difference that there is no need to restrict ourselves to singular why-questions. Using the notion of homogeneous variation sketched in section 3.1, let us say that an admissible contrast-class must be a set of propositions satisfying van Fraassen's conditions and subject to the further demand that all the propositions exhibit homogeneous variation. An ideal why-question acceptable relative to K is a triple $<P,X,R>$ where P is expressed by some member of K, X is an admissible contrast-class, and R obtains between a sequence of propositions A and $<P,X>$ just in case A is expressed by a derivation in $E(K)$ whose conclusion expresses the conjunction of P and the negations of the remaining members of X. An actual why-question acceptable relative to K is a triple $<P,X,R>$ where P, X must satisfy the same conditions as before and R holds between A and $<P,X>$ just in case A is a subsequence of a sequence of propositions expressed by a derivation in $E(K)$ whose conclusion expresses the conjunction of P and the negations of the remaining members of X.

To say that a why-question is acceptable relative to K is not to imply that other why-questions should be rejected in a context where the background beliefs are the members of K. For we ought to allow for the possibility that a why-question might be answered by producing a derivation among whose premises is some proposition (or propositions) that is not expressed by any statement in K but which would be rationally accepted by those who believe the members of K. It is even possible that why-questions should be answered by derivations instantiating patterns that are not in the basis of $E(K)$ but that would be included in the basis of $E(K^*)$ where K^* would be rationally accepted by those who accept K and who recognize the validity of the derivations in question. We want to allow that

science should make progress by appreciating new possibilities for explanation. So I shall say that why-questions that are included under these possibilities are acceptable relative to *K in the extended sense*.

4.5 Explanatory Unification and Causal Dependence

So far, I have developed the view of explanation as unification in a rather abstract way, and it may seem rather ethereal by contrast with the causal approach. One obvious attraction of the latter is that it seems to give a compelling diagnosis of the asymmetry and irrelevance problems. As we shall see, the resolution of these problems is a major obstacle that the unification view must try to overcome, but, before we proceed any further, it is worth relating the claims of the unification approach to the idea that the explanatory asymmetries signal causal asymmetries. *For this is not something that a proponent of the unification view ought to deny*. What is distinctive about the unification view is that it proposes to ground causal claims in claims about explanatory dependency rather than vice versa. So we account for the intuition that appeals to shadows do not explain the heights of towers because shadow lengths are causally dependent on tower heights, by suggesting that our view of causal dependency, in this and kindred cases, stems from an appreciation of the explanatory ordering of our beliefs.

How can this be? Surely few people have any explicit knowledge of the explanatory patterns immanent in the practice of scientists, and fewer still could articulate the factors that contribute to unifying power. So the idea that any one individual justifies the causal judgments that he/she makes by recognizing the patterns of argument that best unify his/her beliefs is clearly absurd. However, in claiming that causal dependence is grounded in explanatory dependence, the champions of the unification approach need commit themselves to no such implausible story. Our everyday causal knowledge is gained by absorbing the lore of our community. The scientific tradition has articulated some general patterns of derivation – sometimes explicitly considering how the phenomena within a domain could be unified, sometimes only under the tacit guidance of the methodological directive to use the minimum of patterns in generating the maximum of conclusions. Derivations that accord with these patterns become accepted as explanatory, and the phenomena described in their conclusions are viewed as objectively dependent on the phenomena described in their premises. So there passes into our common ways of thinking, and our common ways of talking, a view of the ordering of phenomena, and this picture of how phenomena are ordered is expressed, often though not invariably, in our recognition of causal dependencies. Thus the picture advanced by the unification approach shows the concept of causal dependence as derivative from that of explanatory dependence, but it does not promote the dubious idea that each of us gains explicit knowledge of causal dependencies through recognition of the structure of the explanatory store.

My claim that the structure of the explanatory store gives rise to views of objective dependence which are *often though not invariably* expressed in our causal discourse signals the ability of the unification approach to accommodate some of the difficulties that we noted at the end of section 3.2. For even in areas of investigation where causal concepts do not apply—such as mathematics—we can make sense of the view that there are patterns of derivation that can be applied again and again to generate a variety of conclusions. Moreover, the unification criterion seems to fit very well with the examples in which explanatory asymmetries occur in mathematics. Derivations of theorems in real analysis that start from premises about the properties of the real numbers instantiate patterns of derivation that can be used to yield theorems that are unobtainable if we employ patterns that appeal to geometrical properties. Similarly, the standard set of axioms for group theory covers both the finite and the infinite groups, so that we can provide derivations of the major theorems that have a common pattern, while the alternative set of axioms for the theory of finite groups would give rise to a less unified treatment in which different patterns would be employed in the finite and in the infinite cases. Lastly, what Lagrange seems to have aimed for is the incorporation of the scattered methods for solving equations within a general pattern, and this was achieved first in his pioneering memoir and later, with greater generality, in the work of Galois.

The fact that the unification approach provides an account of explanation, and explanatory asymmetries, in mathematics stands to its credit, but this may be viewed as too small a benefit to count very heavily in its favor. By contrast, the problem of understanding theoretical explanation is surely of the highest importance. In the rest of this section I shall try to show that the unification approach gives us insight into the ways in which scientific theories yield explanations. Not only will the discussion of particular examples provide concrete illustrations of the notions of pattern and unification. It will also reveal how the unification approach leads to an improved understanding of important metascientific concepts, revising our ideas about theories, laws, and reduction.

4.6 Unification and Theoretical Explanation

I shall begin by recalling an important point made by Thomas Kuhn (1970, 23–51, 181–191) and, in a somewhat different way, by Sylvain Bromberger (1963). When we conceive of scientific theories as sets of statements (preferably finitely axiomatized) then we naturally think of knowing a scientific theory as knowing the statements—typically knowing the axioms and, perhaps, some important theorems. But, as Kuhn points out, even in those instances where there are prominent statements that can be identified as the core of the theory, statements that are displayed in the texts and accompanied with names—as, for example, Maxwell's equations, Newton's laws, or Schrödinger's equation—it is all too

common for students to know the statements and yet to fail to understand the theory, a failure signaled by their inability to do the exercises at the end of the chapter. Scientific knowledge involves more than knowing the statements. A good account of scientific theories should be able to say what the extra cognitive ingredient is.

I claim that to know a theory involves the internalization of the argument patterns associated with it, and that, in consequence, an adequate philosophical reconstruction of a scientific theory requires us to identify a set of argument patterns as one component of the theory. This is especially obvious when the theory under reconstruction is not associated with any "grand equations" and when reconstructions of it along traditional lines produce a trivialization that is remote from the practice of scientists. I shall consider three examples: two (classical genetics and nèo-Darwinian evolutionary theory) in which traditional philosophical approaches seem to me to have provided very little insight and one (the theory of the chemical bond) that is rarely discussed but that would seem to offer a more ready application of standard methods of analysis.

4.6.1 Classical Genetics

When we read the major papers of the great classical geneticists or when we read the textbooks in which their work is summarized, we find it hard to pick out any laws about the transmission of *all* genes. These documents contain information about the chromosomal arrangement of particular genes in particular organisms, about the effect on the phenotype of particular mutations, about frequencies of recombination, and so forth. In works from the pre-Morgan era we do find two general statements about gene transmission—Mendel's Laws (or "Rules" as they are sometimes called)—but, in the heyday of classical genetics (1910–1953), the writings of classical geneticists are predominantly concerned either with cases to which Mendel's simple laws do not apply or with instances in which they are false. The heterogeneous collection of particular claims advanced about linkage, position effect, epistasis, nondisjunction, and so forth, in *Drosophila, Zea mays, E.coli, Neurospora*, etc. seem more like illustrations of the theory than core principles of it. That, I suggest, is precisely what they are.

Classical genetics is centrally focused on (though by no means confined to) a family of problems about the transmission of traits. I shall call them *pedigree problems*, for they are problems of identifying the expected distributions of traits in cases where there are several generations of organisms related by specified connections of descent. The questions that arise can take any of a number of forms: What is the expected distribution of phenotypes in a particular generation? Why should we expect to get that distribution? What is the probability that a particular phenotype will result from a particular mating?, and so forth. Classical genetics answers such questions by making hypotheses about the relevant genes,

their phenotypic effects, and their distribution among the individuals in the pedigree. Each version of classical genetic theory contains a problem-solving pattern exemplifying this general idea, but the detailed character of the pattern is refined in later versions so that previously recalcitrant cases of the problem can be accommodated.

I shall consider four examples of explanatory schemata that have been employed in genetics in our century. By exhibiting them in detail, I hope to illustrate concretely how explanatory unification of a field works and how changes within a field can show cumulative modification of an underlying pattern. Obviously, picking out four points in a long historical development gives the illusion that the history proceeds in large jumps. Those familiar with the history of genetics in the twentieth century should be able to see how to interpolate and to extrapolate.[19]

[1] **Mendel** *(1900)*

(1) There are two alleles *A*, *a*. *A* is dominant, *a* recessive.

(2) *AA* (and *Aa*) individuals have trait *P*, *aa* individuals have trait *P'*.

(3) The genotypes of the individuals in the pedigree are as follows: i_1 is G_1, i_2 is G_2, . . . , i_N is G_N. {(3) is accompanied by a demonstration that (2) and (3) are consistent with the phenotypic ascriptions in the pedigree.}

(4) For any individual *x* and any alleles *yz* if *x* has *yz* then the probability that *x* will transmit *y* to any one of its offspring is ½.

(5) The expected distribution of progeny genotypes in a cross between i_j and i_k is *D*; the expected distribution of progeny genotypes in a cross . . . {continued for all pairs for which crosses occur}.

(6) The expected distribution of progeny phenotypes in a cross between i_j and i_k is *E*; the expected distribution of progeny phenotypes in a cross . . . {continued for all pairs in which crosses occur}.

Filling Instructions: *A*, *a* are to be replaced with names of alleles, *P*, *P'* are to be replaced with names of phenotypic traits, i_1, i_2, . . . , i_N are to be replaced with names of individuals in the pedigree, G_1, G_2, . . . , G_N are to be replaced with names of allelic combinations (e.g. *AA*, *Aa*, or *aa*), *D* is replaced with an explicit characterization of a function that assigns relative frequencies to genotypes (allelic combinations), and *E* is to be replaced with an explicit characterization of a function that assigns relative frequencies to phenotypes.

Classification: (1), (2), and (3) are premises; the demonstration appended to (3) proceeds by showing that, for each individual *i* in the pedigree, the phenotype assigned to *i* by the conjunction of (2) and (3) is that assigned in the pedigree; (4) is a premise; (5) is obtained from (3) and (4) using the principles of probability; (6) is derived from (5) and (2).

Comments: **Mendel** is limited to one locus, two allele cases with complete dominance. We can express this limitation by pointing out that the pattern above does not have a correct instantiation for examples which do not conform to these conditions. By refining **Mendel**, we produce a more complete schema, one that has correct instantiations in a broader class of cases.

[2] **Refined Mendel** *(1902?–1910?)*

(1) There are n pertinent loci $L_1, \ldots, L_n$. At locus L_i there are m_i alleles $a_{i1}, \ldots, a_{imi}$.

(2) Individuals who are $a_{11}a_{11}a_{21}a_{21} \ldots a_{n1}a_{n1}$ have trait P_1; individuals who are $a_{11}a_{12}a_{21}a_{21} \ldots a_{n1}a_{n1}$ have trait P_2; . . . {Continue through all possible combinations.}

(3) The genotypes of the individuals in the pedigree are as follows: i_1 is G_1, i_2 is $G_2, \ldots, i_N$ is G_N. {Appended to (3) is a demonstration that (2) and (3) are consistent with the phenotypic ascriptions given in the pedigree.}

(4) For any individual x and for any alleles y, z, if x has yz then the probability that a particular one of x's offspring will have y is ½.

(5) The transmission of genes at different loci is probabilistically independent.

(6) The expected distribution of progeny genotypes in a cross between i_j and i_k is D; the expected distribution of progeny genotypes in a cross . . . {continued for all pairs in the pedigree for which crosses occur}.

(7) The expected distribution of progeny phenotypes in a cross between i_j and i_k is E; the expected distribution of progeny phenotypes in a cross . . . {continued for all pairs in the pedigree for which crosses occur}.

Filling Instructions: very similar to those for **Mendel** (details are left to the reader).

Classification: (1), . . . , (5) are premises; (6) is derived from (3), (4), and (5) using principles of probability; (7) is derived from (2) and (6).

Comments: **Refined Mendel** can cope with examples in which a phenotypic trait has a complex genetic basis; it is freed from the limitation to complete dominance and recessiveness and it can allow for epistasis. But **Refined Mendel** does not take account of linkage and recombination. The next step is to build these in.

[3] **Morgan** *(1910–1920)*

(1)–(4) As for **Refined Mendel**.

(5) The linkage relations among the loci are given by the equations $Prob(L_i, L_j) = p_{ij}$. $Prob(L_i, L_j)$ is the probability that the alleles at L_i, L_j on the same chromosome will be transmitted together (if L_i, L_j are loci on the same chromosome pair) and is the probability that arbitrarily selected alleles at L_i, L_j will be transmitted

together (otherwise). If L_i, L_j are loci on the same chromosome pair, then $0.5 \leq p_{ij} \leq 1$. If L_i, L_j are on different chromosome pairs, then p_{ij} is 0.5.

(6) and (7). As for **Refined Mendel**.

Filling Instructions and *Classification*: As for **Refined Mendel**.

Comments: Whereas in **Refined Mendel**, we had two general laws about the transmission of genes—namely (4) and (5)—one of these has given way to a schematic sentence which can be differently instantiated in different situations (depending on underlying cytogological details). Within Classical Genetics, **Morgan** is further refined after 1920 to allow for nondisjunction, duplication, unequal crossing over, segregation distortion, cytoplasmic inheritance, and meiotic drive. I shall leave to the reader the task of showing how all but one of these phenomena are accommodated by modifying **Morgan**. The exception—no more important than the others for the history of classical genetics, but especially interesting for the present study—is meiotic drive. Consideration of meiotic drive requires the abandonment of the *other* general law about the transmission of genes that has so far figured in our schemata, (4), and its replacement by a schematic sentence that can be instantiated differently in different cases. I shall introduce this modification simultaneously with developing another, namely the embedding of classical genetics within molecular biology.

[4] **Watson-Crick**

(1) There are n loci $L_1, \ldots, L_n$. At locus L_i there are m_i alleles $a_{i1}, \ldots, a_{imi}$.

(2) (a) The DNA sequence of a_{11} is *XYUV* . . . , the DNA sequence of . . . {continue through all alleles}.

(b) Details of transcription, post-transcriptional modification, and translation for the alleles in question.

(c) The polypeptides produced by $a_{11}a_{11}a_{21}a_{21} \ldots a_{n1}a_{n1}$ individuals are $M_1, \ldots, M_k$, the polypeptides produced by . . . {continue for all allelic combinations}.

(d) Details of cell biology and embryology for the organisms in question.

(e) Individuals who are $a_{11}a_{11}a_{21}a_{21} \ldots a_{n1}a_{n1}$ have phenotype P_1, individuals who are . . . {continue through all possible combinations}.

(3) The genotypes of the individuals in the pedigree are as follows: i_1 is $G_1, \ldots, i_N$ is G_N. {Appended to (3) is a demonstration that (2e) and (3) are consistent with the phenotypic ascriptions given in the pedigree.}

(4) If an individual x has $a_{11}a_{12}$ at locus L_1 then the probability that a particular offspring of x will receive a_{11} is q_{112}, if an individual x has . . . {continue through all heterozygous combinations}.

(5)–(7).As for **Morgan**.

Filling Instructions: As for **Morgan**, with the further condition that $X,Y,U,V, \ldots$ in (2a) are to be replaced with names of bases (Adenine, Cytosine, Guanine, Thymine) and that the M_i in (2c) are to be replaced with names of polypeptides.

Classification: (2c) follows from (2a) and (2b); (2e) is derived from (2c) and (2d). Otherwise, as for **Morgan**.

Comments: The replacement of (4) with a schematic sentence allows us to accommodate cases of meiotic drive, for these will be represented as situations in which one of the q_{ijk} is different from 0.5. The contribution of molecular biology consists in the *extension* of **Morgan** through the derivation of what was previously a premise (2), which now appears as a conclusion (2e). There are several instances in which the molecular derivation can proceed as far as (2c), but the information about cell physiology and embryology (2d) is always too sparse to permit us to make a *complete* derivation of (2e). One of the closest approximations is furnished by studies of human sickle-cell anemia (and of the molecular structure of the genes for globin chains), and such instances show how the derivation would *ideally* be carried out for phenomena that are developmentally more complex.[20]

Summary. These examples of four main explanatory patterns are intended to show concretely how explanation-seeking questions about the transmission of traits through pedigrees – specifically questions about why we should expect to find particular distributions in specified generations – are addressed within genetic theory. As I have already noted, the successive articulations of the initial pattern, **Mendel**, appears to be cumulative, and, as we proceed, it becomes hard to identify any general laws about the transmission of genes. I shall return to the significance of this case for our understanding of various important metascientific concepts after we have looked more briefly at two other examples.

4.6.2 Darwinian Evolutionary Theory

Darwin's theory of evolution by natural selection addresses a number of general questions about the characteristics of life. These questions include problems of biogeography, of the relationships among organisms (past and present), and of the prevalence of characteristics in species or in higher taxa. As I have argued elsewhere (Kitcher 1985a), Darwin's principal achievement consisted in his bringing these questions within the scope of biology, by showing, in outline, how they might be answered in a unified way.

In its most general form, Darwin's proposal is to make history central to the understanding of biological phenomena. We explain the distribution of organisms in a particular group – the Galapagos finches or the ring-tailed lemurs, for example – by tracing a history of descent with modification that charts the movements of the organisms in the lineage that terminates with the group in which we are interested. In similar fashion, history can be made relevant to the explanation

of relationships among organisms or of the presence of prevalent traits. Historical explanations of this general type divide into two subsidiary classes: in some cases there is an attempt to give a causal account of the modifications in the lineage; in other instances we simply record the modifications without any attempt to identify their causes. In providing (or sketching) explanations in the latter class, Darwin draws only on the less controversial part of his theory, contending that organisms have evolutionary histories but not committing himself to claims about the agents of evolutionary change. When he and his descendants attempt to give the more ambitious historical explanations that specify causes of particular evolutionary changes – most prominently in their efforts to understand why particular traits are prevalent in particular groups – then they venture into areas that were highly controversial long after the idea of evolution was broadly accepted and that still excite (different) debates today.[21]

I shall illustrate the explanatory structure of Darwinian evolutionary theory by displaying two patterns, one from the less ambitious class and one that uses the key Darwinian idea of natural selection. Consider first the explanation of homologous characteristics in related groups. Here the question that confronts us takes the form "Why do G and G^* share the common property P?" In general, questions of this type may be answered by instantiating any of several different patterns – for common traits may occur as a result of parallelism or convergence, instead of being homologies. However, Darwin proposes that many questions of the form can be answered by discovering true premises that instantiate schemata in the following pattern.

Homology

(1) G and G^* descend from a common ancestral species S.

(2) Almost all organisms in S had property P.

(3) P was stable in the lineage leading from S to G: that is, if S was ancestral to S_n and S_n immediately ancestral to S_{n+1} and S_{n+1} ancestral to G, then if P was prevalent in S_n almost all members of S_{n+1} were the offspring of parents, both of whom had P.

(4) P was stable in the lineage leading from S to G^*.

(5) P is heritable: that is, almost all offspring of parents both of whom have P will have P.

(6) Almost all members of G have P and almost all members of G^* have P.

Filling Instructions: P is to be replaced by the name of a trait, G and G^* by the names of groups of organisms (populations, species, genera, higher taxa), S by the name of a species.

Classification: (1)–(5) are premises; (6) is derived from (1)–(5) using mathematical induction on the lineages.[22]

Let us now turn to the primary Darwinian pattern, the pattern underlying *many*

explanations of the presence of traits in groups of organisms. Here, the explanation-seeking question with which we are concerned is "Why do almost all the organisms in *G* have *P*?"

Simple Selection

(1) The organisms in *G* are descendants of the members of an ancestral population *G** who inhabited an environment *E*.

(2) Among the members of *G** there was variation with respect to *T*: some members of *G** had *P*, others had *P#*, *P##*, . . .

(3) Having *P* enables an organism in *E* to obtain a complex of benefits and disadvantages *C*, making an expected contribution to its reproductive success *w(C)*; having *P#* enables an organism to obtain a complex of benefits and disadvantages *C#*, making an expected contribution to its reproductive success *w(C#)*; . . . [continued for *P##* and all other variant forms of *T* present in *G**]. $w(C) > w(C\#)$, $w(C) > w(C\#\#)$, etc.

(4) For any properties P_1, P_2, if $w(P_1) > w(P_2)$ then the average number of offspring of organisms with P_1 that survive to maturity is greater than the average number of offspring of organisms with P_2 that survive to maturity.

(5) All the properties *P*, *P#*, *P##*, . . . are heritable.

(6) No new variants of *T* arise in the lineage leading from *G** to *G* (i.e., the only variation with respect to *T* comprises the properties *P*, *P#*, *P##*, . . . already present in *G**). All the organisms in this lineage live in *E*.

(7) In each generation of the lineage leading from *G** to *G* the relative frequency of organisms with *P* increases.

(8) The number of generations in the lineage leading from *G** to *G* is sufficiently large for the increases in the relative frequency of *P* to accumulate to a total relative frequency of 1.

(9) All members of *G* have *P*.

Filling Instructions: *T* is to be replaced by the name of a determinable trait (a "character-type"), *P*, *P#*, *P##*, . . . are to be replaced with names of determinate forms of the trait, *G** with the name of an ancestral species, *E* with a characterization of the environment in which members of *G** lived, *C*, *C#*, and so forth are to be replaced with specifications of sets of traits, and *w(C)*, *w(C#)* are replaced with non-negative numbers.

Classification: (1)–(6), (8) are premises; (7) is derived from (1)–(6); (9) is derived from (7) and (8).

Comments: **Simple Selection** is a pattern that can be attributed to Darwin and his early followers—in the sense that although the formalism that I have offered is absent from their writings, it is nonetheless implicit in the explanations of the prevalence of traits that are found in the *Origin* and other Darwinian texts. How-

ever, **Simple Selection** does not cover all of the examples studied by early Darwinians. In some instances (as with Darwin's explanation of the role of selection in the evolution of eyes) the pattern employed seems to involve a number of iterations of simple selection. Moreover, there are Darwinian explanations – notably explanations of quantitative characteristics – that relax (6), allowing that as an advantageous variant is increasing in frequency, a yet more successful trait may arise in the population. A pattern capturing this idea can be presented relatively easily (call it **Directional Selection**). Both iterations of **Simple Selection** and **Directional Selection** can be subjected to a further, gradualistic, constraint, in that we can demand that the properties that increase in frequency at successive stages in the derivation should form a "continuous" sequence (in the obvious, but imprecise, sense that the differences among adjacent members of the sequence are small).

There is no reason to insist that all Darwinian answers to questions of the form "Why is *P* prevalent in *G*?" should be instantiations of **Simple Selection**, iterations of **Simple Selection**, or **Directional Selection**. Darwin's insightful appeals to "Correlation and Balance" can be accommodated by recognizing a pattern of **Correlated Selection**, in which the increase in frequency of a characteristic *P* is explained by using one of the selectionist patterns to show that another trait *Q* will increase in frequency, and using a premise asserting the correlation of *P* and *Q* to derive the conclusion that *P* increases in frequency.

When we turn from Darwin to contemporary evolutionary theory, we observe the same kind of explanatory extension that we have already seen in the case of classical genetics. In some ways, the extension is more impressive in the present case, for there are several places in **Simple Selection** (and in the other selectionist patterns) in which schematic premises can be derived as the conclusions of detailed schematic arguments. So, for example, instances of (5) can be derived from specifications of the genetic basis of the trait under study, and the use of population genetics can yield a more precise version of (4) and the derivation of a conclusion about rates of increase in the relative frequency of *P* that will imply (7). Using population genetics and a premise specifying the number of generations in the lineage between *G** and *G*, it is also possible to obtain (8). Finally, the use of ecological models (such as optimality models or game-theoretic analyses) enables us to derive precise claims that imply (3).[23] Thus **Simple Selection** becomes embedded in a much larger argument pattern that can, in principle and occasionally in practice, be instantiated to give neo-Darwinian explanations of the prevalence of traits.

4.6.3 The Theory of the Chemical Bond

The examples that I have been considering involve theories that have proved difficult to reconstruct using the traditional kinds of philosophical formalism. The last case to be discussed focuses on a theory that one might initially regard as far

more amenable to those formalisms. I want to suggest a different way of looking at the structure of a part of chemistry as it has developed from the early nineteenth century to the mid-twentieth century.

One leading question of post-Daltonian chemistry concerned the ratios of the weights of substances that form compounds together. Within nineteenth-century chemistry, we can discern the following simple pattern for answering questions of the form "Why does one of the compounds between X and Y always contain X and Y in the weight ratio $m{:}n$?":

Dalton

(1) There is a compound Z between X and Y that has the atomic formula X_pY_q.

(2) The atomic weight of X is x; the atomic weight of Y is y.

(3) The weight ratio of X to Y in Z is $px{:}qy$ $(= m{:}n)$.

Filling Instructions: X,Y,Z are replaced by names of chemical substances; p,q are replaced by natural numerals; x,y are replaced by names of real numbers.

Classification: (1) and (2) are premises, (3) is derived from (1) and (2).

Dalton is elementary—although it was, of course, instantiated in many different ways during the early years of the nineteenth century by chemists who had very different ideas about the formulas of common compounds! What makes it important for our purposes is the way in which **Dalton** was extended by subsequent work.

The first step in the extension was the introduction of the concept of valence and rules for assigning valences that enabled chemists to derive conclusions about which formulas characterized possible compounds between substances. Thus instances of (1) could be derived from premises ascribing valences to the substances under study (i.e., to X and Y), from premises stating the constraints on formulas for possible compounds in terms of valence relationships, and from the principle that all compounds corresponding to formulas that meet all the constraints can be formed. At this first stage, the attributions of valence are unexplained and there is no understanding of why the constraints hold. However, the original explanations of weight relationships in compounds are deepened, by showing regularities in the formulas underlying compounds.

The second stage consists in the introduction of a shell model of the atom to explain the hitherto mysterious results about valences. From premises attributing shell structure to atoms, together with principles about ionic and covalent bonding, it is now possible to provide derivations of instances of (1). These derivations provide a deeper understanding of the conclusions than was given by the simple invocation of the concept of valence because they show us *in a unified way* how the apparently arbitrary valence rules are generated. Moreover, the appeal to the model of the atom enables us to derive instances of (2) from premises that characterize the composition of atoms in terms of protons, neutrons, and electrons.

Finally, the derivations given at the second stage can be embedded within quantum mechanical descriptions of atoms and the shell structures and possibilities of bond-formation revealed as consequences of the stability of quantum-mechanical systems. Although this is only mathematically tractable in the simplest examples, it does reveal the ideal possibility of a further extension of our explanatory derivations.

4.6.4 Conclusions from the Examples

My three examples are intended to illustrate the notion of pattern that is employed in my account of explanation, and to show that it is not so remote from the practice of science as the abstract description of the earlier parts of this section might suggest. But I hope they do more. Specifically, I would like to suggest that they shed light on some important metascientific concepts.

Laws. As we have seen, Hempel's account of scientific explanation faced serious difficulties in giving a characterization of scientific laws. One way to try to understand the notion of a scientific law is to turn the Hempelian account on its head. So we can suggest that the statements accepted as laws at a given stage in the development of science (recall that our approach has been focusing on *acceptable* explanations not on *correct* explanations) are the universal premises that occur in explanatory derivations. Many of these will be "mini-laws": such as the statement that specific genotypes regularly give rise in particular environments to certain phenotypes or the statement that sodium and chlorine combine in a one-one ratio. My earlier claim that some sciences may not be identifiable by concentrating on a few grand equations may now be made more precise by proposing that a maxi-law is a nonschematic universal premise that occurs in an explanatory pattern. As the example of genetics shows, there were stages in classical genetics at which it seemed that there were maxi-laws in genetics—general principles about the transmission of genes—namely Mendel's laws. But all our examples point to the possibility that sciences may have no maxi-laws, and that their generality may consist in the patterns of derivation that they bring to the explanation of the phenomena.

Reduction. The classical way of thinking about reduction is to consider the derivability of laws of the reduced theory from laws of the reducing theory. If theories are viewed as constituted by the patterns of derivation they put forward (by the ordering of the phenomena that they propose) then the important notion becomes that of *explanation extension*. As the examples reveal, explanation extension can go forward even when some of the concepts of the extended theory cannot be formulated in terms of the concepts of the extending theory. Hence, even when the conditions demanded by reductionists cannot be met, we can still make clear the relations among successive theories, and so capture the idea of an accumulation of knowledge which often seems to make reductionism attractive.

So I propose that the outmoded concept of reduction, which is tied to an inadequate account of scientific theories, should be replaced with the notion of explanation extension, and disputes about the virtues of reductionism reformulated accordingly.

Unification. When the view that explanation is unification is intially presented, I think that it strikes many people as invoking a rather ethereal ideal. However, in the examples I have discussed, we do find that a single pattern of derivation (or several closely related patterns of derivation) is (are) used again and again to derive a variety of conclusions. Thus I take the examples to provide *prima facie* support for the view that unification is important to explanation and that unification works in the way that I have suggested.

With two approaches to explanation now before us, we are in a position to see how they fare with respect to the problems adduced in section 1. In the next section I shall consider whether the unification approach can handle statistical explanation.

5. A Defense of Deductive Chauvinism

One obvious deficiency of the systematization account, as I have so far presented it, is that nothing has been done to accommodate the two kinds of difficulty that surround the issue of probabilistic explanation. First, while there is no bar in principle to the use of nondeductive arguments in the systematization of our beliefs, so that we can make sense of the notion of a general argument pattern even in cases where the instantiations are not deductive, the task of comparing the unifying power of different systematizations looks even more formidable if nondeductive arguments are considered. Second, and perhaps more obvious, is the objection that the systematization account is fundamentally wedded to the old Hempelian idea that to explain is to derive, so that it will be forced to adopt some version of the high-probability requirement, and thereby prove vulnerable to the criticisms leveled by Jeffrey, Salmon, and Coffa against Hempel's account of inductive-statistical explanation. I believe that both problems can be sidestepped by a simple and radical step. The explanatory store contains only deductive arguments. *In a certain sense*, all explanation is deductive.

Following Salmon (who took the label from Coffa), I shall call the thesis that all explanation is deductive *deductive chauvinism*. As I have just noted, deductive chauvinism is logically independent of the systematization account, but a successful elaboration of deductive chauvinism would enable the systematization account to circumvent what appear to be serious problems about probabilistic and statistical explanation. My goal in the present section is to elaborate deductive chauvinism, and show that it is not so absurd a view as it might initially appear.

Now there are some kinds of explanation involving probabilities that pose no threat to the deductivist ideal of explanation. When a geneticist explains why the

probability of obtaining a particular phenotype from a specified cross takes on a certain value, what is given is a deductive argument in which the identification of the probability is derived from a set of premises including some claims about the organisms under study and the mathematical principles of probability (see 4.6.1). It is no accident that such arguments are solicited in student exercises in applied probability theory. To revert to Hempelian terminology, if all explanation were *D-S* explanation, there would have been no need to confront the problems that Hempel faced in developing his *I-S* account.

Trouble arises because there are areas of science—most obviously in applied science—where particular occurrences or states of affairs are explained by appeal to probabilities. The classic examples involve the recovery of patients who have been administered drugs that have known frequencies of success, the committing of crimes by people with a specified age, sex, and socioeconomic background, the occurrence of particular phenotypic traits in organisms whose parents have particular phenotypes, and so forth. Now in all of these cases the deductivist has a relatively straightforward gambit: we treat the probabilistic account as a placeholder for an underlying, unknown, deductive explanation. Thus, if we explain why an individual a has property P by pointing out that a has properties Q, R, S, T and showing that $Pr(P/Q\&R\&S\&T)$ takes some (high) value p then we are treating the high value of p as an indication that there are further, as yet unspecified properties of a, X, Y, Z such that *all* entities that have Q, R, S, T, X, Y, Z have P. The probabilistic argument is pragmatically successful because we have grounds for thinking that it exhibits part of a *deductive basis* for the phenomenon to be explained, where a deductive basis consists of those properties that would be attributed to the object(s) mentioned in the *explanandum* in the premises of a complete deductive explanation. Those grounds consist in the fact that the probability value p is high. Hempel's high-probability requirement returns as a claim about what would provide evidence for thinking that we have identified part of the deductive story.[24]

While I think that this straightforward gambit expresses part of what a deductive chauvinist should say about statistical explanation of individual events, it cannot be a completely satisfactory solution to the problem. Even if there are legitimate hopes that our uses of probability in medical and social contexts are only expressions of ignorance, there is at least one area of theoretical science in which a case can be made for the necessity of appealing to probabilities in explaining individual events and states. Quantum mechanics (QM) seems to be indeterministic. Moreover, an opponent of deductive chauvinism may contend that the indeterminacies of QM ultimately affect the macroscopic phenomena of chemistry, biology, medicine, and social science, so that the dream of deductive explanations in these areas may prove to be quite unrealistic. The point can be further reinforced by suggesting that there may be additional sources of indeterminacy beyond those already recognized in QM. I shall try to meet these challenges in turn.

5.1 The Objection from Quantum Mechanics

One familiar response to QM is the belief (hope, wish) that it will some day be replaced by a physical theory that will enable us to derive descriptions of individual events from underlying theoretical principles and initial and boundary conditions. The version of deductive chauvinism that I want to defend does not adopt any such attitude. It may well turn out that successor theories to QM retain its indeterministic character. Deductive chauvinism should distinguish between two senses in which an explanatory account might be ideal. In one sense, an ideal explanatory account is a deductive derivation (more exactly, a deductive derivation that instantiates a pattern in the explanatory store). In another sense, an ideal explanatory account is the best that the phenomena will allow. Deductive chauvinists should concede that QM (or some other essentially indeterministic theory) might be the best there is, that it might provide ideal explanatory accounts (of individual events) in the second sense, but they should deny that QM provides ideal explanatory accounts (of individual events) in the first sense.

The issues here can be treated more precisely by returning to the account of the pragmatics of explanation advanced in section 2. Those who contend that QM provides explanations of individual events must believe that there are why-questions whose topics describe such events to which QM provides complete explanatory answers. Faced with any purported example, the deductive chauvinist must do one of two things: either show that the complete answer is a deductive derivation or demonstrate that the answer is not complete (in the latter case, it will also be helpful to explain the *illusion* of completeness).

Let us start with a situation that is typical of those in which we might think that the probabilistic machinery is brought to bear on individual events. Imagine that a beam of electrons impinges on a potential barrier. For each electron, the probability that it will be reflected is 0.9, the probability that it will tunnel through is 0.1. Consider two electrons, e_1, and e_2. e_1 is reflected; e_2 tunnels through. Can we explain these events?

The question urgently needs disambiguation. Suppose we begin with standard (van Fraassen) why-questions, and consider two obvious candidates, both of which have as their topic the proposition that e_2 tunneled through. The contrast class of the first question contains both the topic and the proposition that e_2 was reflected; the contrast class of the second contains both the topic and the proposition that e_1 tunneled through. Then, I claim, there is no explanatory answer to either question—that is, there is no proposition that stands in the relation of ideal explanatory relevance to topic and contrast class.

Now QM enables us to write down the Schrödinger equation for the system consisting of an approaching electron and the potential barrier, and, by solving this equation we can demonstrate the probabilities of tunneling through and being reflected. Conjoin as many as you like of the propositions that occur in this deri-

vation or consider the entire derivation. Whatever your choice, you will not have shown why e_2 tunneled through, rather than being reflected, or why e_2, rather than e_1, tunneled through. For what those why-questions ask is for a specification of the differences between electrons that tunnel through and electrons that are reflected, and it is, of course, part of the character of QM that there are no such differences to be found. In a sense, the full derivation from the Schrödinger equation is the best possible answer to the questions – it is the best that nature will allow – but it is not an ideal explanatory account. With respect to these questions, there is no ideal explanatory account. The questions are unanswerable. In response to the questions we can only say "It turned out this way, by chance."

There is a superficial difference when we consider the related questions about e_1. If someone inquires why e_1 was reflected rather than tunneling through, or why e_1 rather than e_2 was reflected, one may be tempted to use the fact that the derivation from the Schrödinger equation assigns a relatively high probability (0.9) to the event of e_1's being reflected to suggest that that derivation supplies at least something of an answer to these questions (or, perhaps, to the first). After all, the derivation does show that it was to be expected that e_1 would be reflected. But I think that the situation is symmetrical. We no more understand why e_1 was reflected than we understand why e_2 tunneled through, and, in each case, our failure of understanding rests on the fact that we cannot isolate any distinctive property that separates those that take one course from those that take the other. The symmetry here underscores my earlier diagnosis of how the Hempelian high-probability requirement might be expected to work. High probabilities are useful because they increase our confidence that we are on the track of a deterministic basis. But, when we know from the start that there is no deterministic basis to be found, the probability values are irrelevant.

If this is correct, then what, if anything, can be made of the idea that QM does advance our understanding of episodes like that of electron tunneling and reflection? The answer, I believe, is that we confuse questions that QM can answer with those that it cannot. There are relatives to (van Fraassen) why-questions in the vicinity, and QM can provide answers to some of them.

Sometimes the form of words "Why *P*?" can mean "How is it possible that *P*?" Typically, when this occurs, the question is posed in a context in which certain propositions are taken for granted. Thus, what the questioner is really asking is "Given that *Q* (which I firmly believe) how can it also be the case that *P*?" Sometimes, under such circumstances, corrective answers are called for: *Q* is false and one corrects the question by pointing that out. By analogy with the treatment of why-questions in section 2.1, we can identify a how-possibly question as an ordered pair of propositions $<P,Q>$ where *P* is the topic and *Q* is the *background presupposition* (in some cases *Q* may be null). A noncorrective answer to the how-possibly question is an argument that shows that *P* and *Q* are consistent.

Now imagine a naïve questioner asking why e_2 tunneled through – say in some

concrete laboratory setting where the potential barrier is due to an observable object and the electron's tunneling through is revealed by a scintillation on a screen. It is quite possible that the questioner's inquiry arises from astonishment that a particle, an electron, can make its way through a solid object. In responding to this person's question, we would try to tease out the background presuppositions and to show that those presuppositions that are not false are mutually consistent. In other words, we view the apparent why-question as expressing a how-possibly-question – given the character of solid objects and particles, how is it possible that the electron could penetrate? – and we might respond by using the derivation from the Schrödinger equation to show that there is a non-zero probability of tunneling through.

There are other questions that QM can answer. The conclusion of the derivation specifies the probability that an impinging electron will be reflected (tunnel through) and thereby answers the question "What is the probability that an impinging electron will be reflected?" The derivation itself provides an ideal answer to the why-question "Why is the probability that an impinging electron will be reflected 0.9?" (where the contrast class includes all the propositions attributing values in [0,1] to the probability). Of course, the why-question answered here is a question about an entire class of events, and we have the derivation of a generalization. But, as we have already seen (see above p. 427), there are many instances in which answering a why-question completely involves referring the phenomenon to be explained to a general class and then showing why a regularity holds within the class.

So deductive chauvinists should pursue a strategy of divide-and-conquer with respect to the claim that QM can be used to explain properties of individual events. There are some why-questions that are unanswerable and related questions that have ideal, deductive answers. The apparent difficulties for deductive chauvinism, and the apparent need for the *I-S* model of explanation (or some surrogate) arises from the fact that these questions can be presented in the same form of words, and can thus easily be confused.

5.2 The Idealization of Macro-Phenomena

But now the second worry seems to arise with increased force. Consider any deductive explanation of a macroscopic event. Given QM, we know that if enough fundamental particles entered highly improbable states, some of the claims made in the premises of the derivations would be false. Hence it would appear that *all* explanation of individual events needs to be replaced by deductive explanations of why certain kinds of events occur with high probabilities. For example, imagine that we are explaining why a flood occurred in a house in Minnesota during the occupants' winter vacation. The explanation might derive a description of the phenomenon from principles about the relative density of ice

and water and about rates of cooling, together with specification of initial and boundary conditions. However, if we think of the house and its environment as a quantum-mechanical system, we recognize that there are highly improbable micro-states that would correspond to macro-states in which there is no flooding of the house (in some of these there is no cooling in the pipes, in some of them the pipes are rejoined once the ice melts, and in some of them the house itself decomposes). So, if acceptable explanations are to involve only principles that we judge to be true, then we ought to specify the probability that there will be a flood and provide a full derivation of the specification of the probability.

I think that examples like this help us understand what occurs in the explanation of individual events. When we explain the behavior of actual objects, the first step is always to achieve an idealized description of those objects. Thus, in the standard explanation of the flood in the Minnesota house, we think of the house and its environment as a certain kind of thermal system, assuming uniformities in temperature gradients, ideal cylindrical pipes, and so forth. The question "Why did this particular object behave in this particular way?" is transformed into the question "Why do ideal objects of this general type exhibit these properties?"[25] We justify the transformation by pointing out that factors we have neglected or have introduced in our idealizing either make definite, but small, differences to the account or are highly unlikely to make any large changes. In effect, we are playing "let's pretend," and giving a deductive account of how things would go in a simpler, cleaner, world (for a view that is similar in some respects, see the "Simulacrum account" offered in Cartwright (1983)). We regard the derivations we give as explaining the actual phenomena, because we can provide justifying arguments for concluding that the actual world is not likely to be signficantly different from the ideal world (the probability that there will be a large difference between the phenomena of the actual world and the phenomena of the pretend world is small).

One of the simplest and most celebrated examples in the history of scientific explanation will make the point clearly. According to popular history, the fusiliers at the Venetian artillery asked Galileo why their guns attained maximum range on a flat plain when they were set at an angle of elevation of 45°. Contemporary students in elementary physics learn how to give the substance of Galileo's answer. Consider the gun as a projector, the cannonball as a point particle, the ambient atmosphere as a vacuum, and the plain as an ideal Euclidean plane. Suppose that the particle is subject only to the force of gravity. *Now* we can represent the components of the resultant motion, for varying angle of projection, and show that maximal range is attained when the angle of elevation is 45°. There is an elegant derivation of the result for an ideal system.

But what about the actual guns and the actual cannonballs? Has their behavior been explained? Yes, but the explanation involves correction. It probably is not true that the fusiliers' guns attained maximum range at 45°. The results probably

varied from gun to gun, day to day, cannonball to cannonball, location to location. However, only someone in the grip of a theory of explanation would complain that the presuppositions of the fusiliers' why-questions were false and that the questions could not therefore be answered. So, we can imagine that Galileo, or his contemporary descendants, append to the ideal derivation some remarks about the way in which the actual world and the ideal world are different: in the actual world there are effects of the internal surface of the barrel, of air resistance, of asymmetries in cannonballs, of local inclinations and depressions in the ground; it is possible that the fusiliers have not detected the ways in which these effects produce deviations from the regularity they claim to have found—perhaps the deviations are even too small to be detected with their measuring instruments. The ideal derivation conjoined with an assessment of the extent to which perturbing actual factors would cause deviations from the idealized conclusions gives complete understanding of the actual events.

Here and elsewhere, the idealization ignores both common perturbations with slight effects (such as air-resistance) and major disruptions with negligible probability. The more advanced contemporary physics student knows that the cannonball is a quantum-mechanical system. In consequence, there are extremely small probabilities that unusual motions of its constituent particles will occur, subverting the claims that are made in the equations of the Galilean account. In principle, we could compute the probability that quantum-mechanical effects could generate a significant deviation from the trajectories described in the idealized description—where the standards for significance are set by the magnitude of the macroscopic perturbations we ignore. What is the probability that some QM effect will be more sizable than the effects of air-resistance (say)? Nobody knows the exact answer, but we can make reasonable estimates of order of magnitude and conclude that the probability is very small indeed. Thus part of the defense of our idealization consists in the specification of small perturbations, and the remainder in showing that the probability of a larger difference is negligible.

So the response to the second charge, the charge that the effects of QM percolate up into macrophenomena and subvert the strategy of giving deductive explanations of individual events is that our macroscopic explanations of individual events involve idealizations of the phenomena, deductive derivations that are exactly true of the ideal systems, and assessments of the probable differences between the ideal and the actual systems. The assessments justify our idealizations, and, simultaneously, show us how the topics of our why-questions might be corrected.

5.3 Further Sources of Indeterminism?

It should now be obvious how to respond to the third complaint, the worry that there are sources of indeterminism beyond those of QM.[26] We can consider three

possibilities. First, there are no such sources, and the account given in response to the first two objections suffices. Second, there are such sources, but, like those of QM, they only give small probabilities of significant deviations in macro-phenomena. In this case, the sources would have to be handled in just the same way as the QM effects on macro-phenomena, and we would explicitly note a further idealization of actual macro-events. Third and last, there are sources of significant deviations in macro-phenomena, so that, for some class of macro-phenomena C there are no deductive accounts of the behavior of systems that can be defended as idealizations of the phenomena in C. If this last possibility were to occur, then, I suggest, our attitude toward the macro-phenomena in C should be just that I have recommended we take toward the basic phenomena of QM (such as electron tunneling). In other words, should the basis of heart disease (for example) turn out to be irreducibly probabilistic, then we should have to admit that we can no more explain why one person rather than another contracts heart disease than we can explain why one electron rather than another tunnels through a barrier. Hence, in all three cases, the strategies developed in response to the first two objections will suffice to defend deductive chauvinism.

5.4 Two Popular Examples

Some of the points I have been making can be underscored by considering two examples that have figured in the recent literature and that have sometimes been considered to devastate deductive chauvinism. Let us begin with a case that is used by Salmon (1984, 86, 88, 109). A breeding experiment on pea plants produces a filial population in which 0.75 of the plants have red blossoms, 0.25 white. Let b_1 be a plant with red blossoms, b_2 a plant with white blossoms. Salmon argues that we can explain *both* why b_1 has red blossoms and why b_2 has white blossoms by pointing out that the flowers came from the filial population. The case is intended to illustrate a symmetry principle: we understand the improbable outcome just as much—or as little—as we understand the probable outcome. Thus, Salmon contends, there can be probabilistic explanations of individual events, even in cases where, relative to the information given in the explanatory answer, the probability of the event explained is low.

I believe that these conclusions are mistaken and that the mistakes are revealing. First, we should note that the answers to the questions "Why does b_1 have red blossoms?" "Why does b_2 have white blossoms?" are "Because b_1 has genotype *RR* or *Rr* (where *R* is dominant with respect to *r* and codes for a molecule ultimately producing red pigment)" and "Because b_2 has genotype *rr*." No appeals to probability enter here, *unless we suppose that there are irreducibly probabilistic factors that enter into the connection between genotype and phenotype.* Thus, if it is impossible to provide a deductive derivation of the explanandum it is because complications owing to QM (or some other source of indeterminism) allow

for organisms that bear the *R* allele to fail to produce red blossoms (in the environment, assumed standard, in which the plants are grown). If probabilities enter into the explanation of these individual occurrences it is for just the kinds of reasons I considered above, and, as I contended, those reasons do not compel us to admit probabilistic explanations of individual events (or states).

But the question "Why does b_1 have red blossoms?" may be aimed deeper. The questioner may want to know not only the genetic basis of the trait in b_1 but also why b_1 has the genotype it does. Here we might think that a probabilistic account was appropriate: b_1 is obtained from a cross between *Rr* heterozygotes, so that the probability that it will be *RR* or *Rr* is 0.75. A moment's reflection makes it plain that this does not explain why b_1 rather than b_2 produces red blossoms. So should we conclude, as Salmon suggests (1984, 109) that explanation is not implicitly contrastive? No. For the explanatory value of the probabilistic account just given – and of the corresponding probabilistic account for b_2 – is that they introduce properties of the organisms involved in the process that will figure in a complete explanation. b_1 has genotype *RR* (assuming that that is its genotype) because the fertilization process that gave rise to its zygote involved the fusion of two *R*-bearing gametes; b_2's zygote was formed from two *r*-bearing gametes. There is apparently a fully deductive derivation of the specifications of the genotypes of b_1 and b_2 from statements that describe the events that culminate in the respective fertilizations. Once again, if there are sources of indeterminism in these events, then they would have to be reckoned with along the lines canvassed earlier in this section.

How, then, does probabilistic explanation work in genetics? Answer: along the lines discussed in section 4.6.1. The derivations that are provided by genetics show why certain distributions of genes and traits are expected. Operating against this background of expectations, those who are concerned with particular populations – or with particular individuals – can (in principle) trace the histories of the passage of genes, and so see how the actual cases diverge from (or coincide) with the expectations. By instantiating schemata like **Mendel** we can explain why certain events have specified probabilities. When we want an explanation of the actual events, we have to trace the details of history – but note that the tracing is done along the lines laid down in the theoretical picture.

The second case I want to consider is that of the hapless mayor. Why did the mayor develop paresis? He alone among the townspeople had previously contracted syphilis. But the chance that an individual syphilitic develops paresis is low. Have we, therefore, as Salmon suggests (1984, 31–32, 51–53) given a probabilistic explanation of an event, and one which, furthermore, assigns a low probability to the event explained?

Distinguish possibilities. One reason that we may think that the recognition of the mayor's syphilis has explanatory value is that we think of it as *part* of a complete deductive explanation of the paresis. Syphilitics who have some (possibly

complex) property X always contract paresis. At the present stage of our knowledge, we can distinguish the mayor from his nonsyphilitic neighbors and answer the question of why he, rather than any of them, contracted paresis. We do not yet know how to distinguish the mayor from his fellow syphilitics who do not contract paresis. To answer the question why it was the mayor rather than any of *them* who got paresis, we would have to know the additional factor X. So the information about the mayor is helpful because it enables us to answer one why-question ("Why was it the mayor rather than his [nonsyphilitic] neighbors?"), but more information will be required to answer another ("Why was it the mayor rather than certain other syphilitics?").

But perhaps there is no additional factor to be found. If so then the situation is just like that of electron tunneling. We cannot explain why the mayor, rather than other syphilitics, contracted paresis, any more than we can explain why this electron tunneled through. However, the statement that the mayor had syphilis may answer a different why-question. Suppose that the why-question has an explicit presupposition: "Given that one of the townspeople contracted paresis, why was it the mayor?" Only syphilitics get paresis, and he was the only syphilitic in town. Notice that, in this case, we can deduce the *explanandum* from the presupposition of the question and the information given in the answer. Under these circumstances, we can vindicate the idea that the statement that the mayor had syphilis is part of an explanation of *something like* his getting paresis—but the explanation is deductive.

This last possibility is interesting because it corresponds to a common strategy of scientific explanation (see Sober 1983). Sometimes we show that a system is in state X by presupposing that it is in one of the states $\{X, Y_1, \ldots, Y_n\}$ and demonstrating that it cannot be in any of the Y_i. In contemporary evolutionary theory, there are numerous examples of a special case: one shows that if the system starts in any of the Y_i then it ends in X. The classic example is that of the sex-ratio. In an evolving population of sexual organisms with variation in the propensity to produce sons and daughters, if the population begins away from a 1:1 ratio (but not so skewed that all the organisms are of one sex!) then selection will bring the population to a 1:1 ratio (Fisher 1931, see Sober 1983 for use of the example). The strategy instantiated here is to answer a question of form "Given Q, why P?" by producing a derivation of $Q \supset P$.

5.5 Explanation and Responsibility

There is one further point to be addressed before the defense of deductive chauvinism is complete. Salmon has argued that our use of probabilistic explanations of individual events is involved in assignments of responsibility. Suppose that Herman contracts cancer and dies. His widow sues the federal government on the grounds that Herman spent a significant part of his military service in a

region used for atomic testing. The government points out that the base rate for Herman's type of cancer in the population is 0.0001 and that the probability of a person contracting cancer, given the period of exposure that Herman had, is 0.02. Most of Herman's army buddies were more fortunate. Arnold, for example, did not contract cancer. Nonetheless, even though there is no factor to which we can point that will distinguish Herman from Arnold, though Herman's getting cancer is, *all things considered*, simply a matter of (bad) luck, nonetheless, we have a strong intuitive conviction that the government has some responsibility for Herman's death, and that Herman's widow has a legitimate case.

Salmon believes that we cannot underwrite this intuition unless we are prepared to endorse probabilistic explanations of events which, *all things considered*, have low probability. I shall assume that the intuition is correct and try to show how it is possible to support it from within the perspective of deductive chauvinism. In exact parallel to examples that we have considered before, I suggest that we cannot answer the question "Why was it Herman, rather than Arnold, who contracted cancer?" However, as usual, there are why-questions in the vicinity that we can answer. Consider the question "Why did Herman have a significantly greater probability of contracting the type of cancer from which he died than do members of the general population?" An incomplete answer to this question is "Because he spent a large part of his military service in a region used for atomic testing." We would complete that answer by introducing details of the conditions, applying principles of atomic physics and employing facts of human physiology. In principle, there is a deductive demonstration that reveals the probability of cancer to be 0.02 for people who share Herman's experience and 0.0001 for those who do not. We can deepen that derivation by replacing the premise stating that Herman spent such-and-such a period in a region of such-and-such a type with a deductive argument that leads from premises about the government's actions to the conclusion that Herman had the kind of exposure that he did. So, ultimately, there are premises that make reference to government actions that figure in an explanation of why Herman's chances of contracting cancer were significantly higher than those of the general population.

I have no good theory of how the existence of an explanation making reference to the actions of X shows that X is responsible for some facet of the *explanandum*. Neither does Salmon. However, part of his case against deductive chauvinism involves the quite plausible idea that our intuition that the government is responsible for Herman's death (at least in part) rests on our seeing that we invoke claims about the government in a probabilistic explanation of Herman's death. Deductive chauvinists should agree with Salmon's (tacit) views about the link between explanation and responsibility. But they should modify the treatment of the example. In one obvious sense there is no explanation of Herman's death. There is, however, an explanation of why Herman was at greater risk than the general population. This explanation involves the actions of the government. Because the actions

of the government are explanatorily relevant to the increased probability of Herman's contracting cancer *and* because Herman actually contracted cancer, the government is at least partly responsible.

So there is a chauvinist analog of Salmon's claims about responsibility. Where Salmon sees the responsibility as stemming from the fact that governmental actions are described in premises of a probabilistic explanation of an individual event, I regard it as issuing from the fact that those actions are described in a deductive explanation of why Herman had a greater chance of contracting the cancer from which he died. To put the difference starkly: Salmon holds that the government is responsible because it (partially) caused Herman's death; I hold the government responsible because it caused Herman to be at greater risk for a harm which (by his bad luck) befell him.

Stephen Stich has reminded me that accounts of responsibility will have to come to terms with cases in which the action of one party is pre-empted by the action of another. If the government puts Herman at greater risk for cancer, but Herman succumbs to a heart attack, then the government is not, of course, responsible for his death. Quite evidently, there will have to be clauses that connect explanation and responsibility in more complex cases—both on Salmon's account and on mine—but I see no reason to suspect that the parallelism that I have outlined in the simplest instances cannot be preserved.

I conclude that deductive chauvinism can defeat the obvious challenges. Since acceptance of deductive chauvinism facilitates the development of the idea that explanation is unification, I shall henceforth assume that all derivations in the explanatory store are deductive.

6. Epistemological Difficulties for the Causal Approach

Let us recapitulate. We have seen how the Hempelian models of explanation face apparently insuperable difficulties. Two alternative approaches to explanation have been considered. One of these, the view that scientific explanation identifies the causes of phenomena, is relatively easy to understand, and its virtues, especially in tackling the asymmetries of explanation, are apparent. The other approach, construing explanation in terms of unification, is initially more difficult to formulate. I have tried to show how it might be developed and how some apparent problems with it may be overcome. The rest of this essay will be concerned with what I see as the principal troubles of each program. Besides the smaller problems canvassed in sections 3.2 and 3.3, the causal approach, on the one hand, faces the large question of how the epistemology of causation is to be developed. The unification approach, on the other hand, must show how the constraints on the explanatory store enable us to debar those problematic derivations that engender the problems of asymmetry and irrelevance. The present section will delve into the difficulties of providing an adequate account of our knowledge of

causes, once the traditional empiricist idea of making the concept of causation dependent on such notions as law or explanation has been renounced. Subsequent sections will complete my apology for the unification church, by trying to show that *its* central problem is not so recalcitrant as one might fear.

6.1 Hume's Ghost

Sometimes in reading the philosophy of science literature of the 1940s, 1950s, and 1960s, one has the distinct impression that the authors have been spooked by Hume. Causal concepts are viewed with suspicion, if not dread, apparently because they have been placed on the list of empiricistically forbidden notions. The sense that a whiff of sulphur would accompany certain appeals – appeals that would make life so much easier if they only could be made – makes heresy enticing. So, in the wake of logical empiricism, many philosophers of science have made free use of causal concepts, perhaps seeing themselves as shaking off ghostly chains that had seemed to bind their predecessors. Unfortunately, while the mere invocation of Hume's name is not enough to show that such uses are sinful, there are deeper reasons for worrying about causal concepts than a desire to keep one's empiricist conscience pure. We can discover those reasons by returning to Hume's problem of causation in its most general form.

The desire to analyze causation stems from the apparent difficulty of justifying causal judgments. Some of the causal claims that we make are justified. But how does the justification work? Once we have been educated in the causal lore of an ongoing field of science, then it is easy to claim that we simply observe causal relations. But there is very little plausibility to the idea of observing causation, when the observer is a neophyte, say a child. Thus there arises the conviction that we come to make justified causal judgments by observing that certain conditions obtain, and (initially at least) inferring causal claims from the premises that record our observations. Hence we arrive at the project of giving necessary and sufficient conditions for the obtaining of causal relations, formulating those conditions in ways that will dissolve the epistemological mysteries surrounding causation by deploying only concepts whose satisfaction is observationally ascertainable.

With the demise of statistical accounts of causal relations in the late 1970s (see, for example, Cartwright 1979), few of those who want to deploy the concept of causation in analyzing explanation have taken a public stand on the cogency of the line of reasoning rehearsed in the last paragraph. If causal knowledge is observational knowledge, then the apparently implausible implications of that position should be addressed. But, if causal knowledge is inferential knowledge, then we are owed an account of the observational conditions on which causal justifications depend.[27] Yet explicit accounts of causal knowledge are hard to find. Salmon's recent attempt (in his 1984) is a notable exception, and I shall try to show how

hard the epistemological problem is by reference to his careful and sensitive discussion.

6.2 Causal Processes and Causal Interactions

Salmon is forthright in his acknowledgment that he needs to "furnish an analysis of the concept of causality or its subsidiary notions," and he views it as an issue of "intellectual integrity" to "face Hume's incisive critique of causal relations and come to terms with the profound problems he raised" (1984, 135). However, he thinks that traditional empiricist approaches to causation have handicapped themselves by starting with the wrong causal concepts. Instead, he proposes that we can formulate causal claims so as to make clear how we know them if we focus on causal processes and causal interactions rather than on the causal connections among events. Ultimately, the project of understanding the causal structure of the world can be regarded as an attempt to specify the relation that must obtain between two events (conceived as spacetime points) just in case the earlier was a causal factor in the occurrence of the later. The simplest type of case is that in which the two points are linked by a causal process that connects an interaction at the earlier point (c) with an interaction at the later point (e). A paradigm will be the breaking of a window by a ball that was struck by a baseball bat. Here the event e is the breaking of the window, the event c the striking of the ball, and the linking process is the flight of the ball. In such cases there are no intermediate interactions (or, at any rate, none that are relevant to the occurrence of e) so that the causal connection seems decomposable into a pair of interactions and a causal process that (spatio-temporally) links them. Inspired by the paradigm, we might claim that at least the central cases of causal connection can be understood, if we can only analyze the notions of causal interaction and causal process.

As a first formulation, let us say that two events are causally related just in case they are causally connectible, where causal connectibility obtains just in case there is a continuous path through spacetime that is a causal process and that terminates in causal interactions at both ends. Now it seems to follow that we have immediately eliminated some examples that some would count as instances of causal connection: eighteenth-century claims about action-at-a-distance and twentieth-century proposals about the possibility of time-travel (see, for example, Lewis 1976) both appear problematic. For the moment, I shall accept the elimination. We can then pose Salmon's problem as follows. We want to state conditions on causal connection that will pick out the causal paths and the causal interactions, and the notions to which we appeal in our specifications must be acceptable to an empiricist. The last constraint does not involve swearing to disavow some index of banned concepts drawn up by the empiricist establishment. The demand is that we do not introduce conditions when we are unable to explain how we can know whether or not they obtain.

Begin with the idea of a connecting causal process. We have a manifold of spacetime points, and, for any two distinct points in the manifold, a large number of paths connecting them. The task is to filter out the paths that are not causal processes. Evidently there is a fairly large collection that are prime candidates for elimination (waiving worries about the possibility of certain kinds of time-travel): suppose that e is not in the future light-cone of c and conversely; then, we might suggest, there is no causal process connecting c to e. Similarly, any path that contains a pair of points such that neither belongs to the future light-cone of the other can be excluded. But this still leaves us with a lot.

Intuitively, there are three kinds of continuous paths through spacetime. There are causal processes, along which information can be transmitted (see Salmon 1984, 141); there are pseudoprocesses (Salmon's paradigms are shadows and moving spots of light on walls), which are incapable of transmitting information, but are relatively respectable; and there are gerrymandered paths, spatio-temporal junk, which do not have enough integrity to lead us to dignify them as processes – one simple, and relatively mild, example would be a segment of the worldline of a mosquito just prior to its biting me and a segment of my worldline immediately following the biting. The last category comprises paths that are continuous but that lack any stability of properties along them. To say just what this amounts to is tricky, but we could succeed in Salmon's project without distinguishing the pseudoprocesses from the spatio-temporal junk. It will be enough if we can separate the genuine causal processes from the rest.

Salmon attacks the problem of distinguishing causal processes by developing ideas of Reichenbach's. "A causal process is capable of transmitting a mark; a pseudo-process is not" (1984, 142). Combining three of Salmon's principles, (MT 148; ST 154; PCI 155), we can define a causal process in terms of epistemologically prior notions as follows:

> (CP) P is a causal process iff there are spacetime points c, e, such that P links c and e and it is possible that there should be a modification of P (modifying a characteristic that would otherwise have remained uniform) produced at c by means of a single local interaction and that the modified characteristic should occur at all subsequent points from c to e without any subsequent interaction.

(CP) embodies the idea that causal processes are processes that transmit information because they are *markable*. The modality is obviously needed because not all causal processes are actually marked (think of a universe in which there is a single light signal propagated *in vacuo*). The clause banning subsequent interaction is required because any continuous spatio-temporal path could "transmit a mark" if we were allowed continuous interactions that would impose the marking property upon it at every point.

6.2.1 Some Problems about Processes

Plainly (CP) characterizes the notion of causal process by presupposing the concept of causal interaction. Let us assume that we have the concept of causal interaction in hand, and consider whether the definition of causal process does what it is supposed to do. There seem to be at least four different kinds of problems.

(a) *The problem of pseudomarks.* It is possible for pseudoprocesses to transmit pseudomarks in just the way that causal processes transmit genuine marks. Imagine that a car grazes a stone wall and becomes scratched. The car transmits the scratch: that is, its worldline after the scratch consists of temporal segments that have the property of bearing a scratch. But the shadow transmits the property of being the shadow of a scratched car: its worldline, after the crucial moment, consists of temporal segments that have the property of being a shadow(-stage) of a scratched car.

(b) *The problem of derivative marks.* Suppose that a child traveling in the car puts an arm out the window and holds up a flag. The child's action produces a modification in the shape of the shadow. The modification persists without any further interaction. Provided that the arm is not retracted, the shape of the shadow will continue to be different from what it would otherwise have been.

(c) *The problem of no further interactions.* As I noted above, the clause debarring further interactions is needed to exclude the possibility that pseudoprocesses could "transmit marks" in situations where constant intervention produced the same modification along the spacetime path. But virtually all (all?) *actual* processes are always interacting with other processes. Indeed, there are some causal processes—organisms, for example—for which there could not be any transmission of marks unless there were further interactions. In such cases, the clause that proscribes further interactions will have to be reformulated very carefully. Consider an organism that transmits a mark (a superficial smudge, say). Its actual worldline intersects many other causal processes. We want to say, of course, that those intersecting processes are irrelevant to the persistence of the mark. Even if those particular interactions had been absent the modification would have persisted provided that some interactions of the same general kind had been present. But how do we specify what counts as "the same general kind" here? Some interactions (forceful dowsings, for example) would have removed the smudge. So it appears that if we are to apply (CP) to actual macroscopic cases, we shall need a principled distinction between "relevant" and "irrelevant" interactions—where the concept of relevance is itself a causal notion.

(d) *The problem of fortuitous maintenance.* We can arrange for pseudoprocesses to transmit marks by exploiting the idea that some processes do not

require further interventions to keep them going. Imagine that a vehicle equipped with skis is sliding on an ice rink and casting a shadow. A projectile is thrown in such a way that it lands at the edge of the shadow with a horizontal component of velocity equal to that of the shadow of the vehicle. Because the projectile lies across the edge there is an immediate distortion of the shadow shape. Moreover, the distortion persists because the projectile retains its position relative to the vehicle (and to its shadow).

As we shall discover, not all of these difficulties are equally fundamental. However, separately and in combination, they seem to raise serious troubles for Salmon's account of causation—and some of them seem to me to point to deep problems in the epistemology of causation. Let us begin with (a) and (b) where there appears, at first glance, to be an easy solution.

6.2.2 Troubles with Interactions

The obvious suggestion about (a) and (b) is that, in these examples, there is no interaction with the process that is to be "marked." When the shadow takes on the property of being the shadow of a marked car it (the shadow) does not interact with anything—at least it doesn't interact with anything *relevant*. Similarly, when the child thrusts an arm out the window, there is no interaction—at least no *direct* interaction—between the child's arm and the shadow. So we might try to rescue (CP) from examples (a) and (b) by declaring that the alleged markings are pathological and that the pathology consists in the fact that the supposed marking process does not interact with the process which is supposed to be marked.

Tempting as it may seem, this straightforward response is problematic. Recall the point that underlies (c): there are numerous intersections of spacetime paths, and, if the response of the last paragraph is to succeed, then we need to know why none of these intersections can count as a marking process. Specifically, in case (a) there is some intersection between the worldline of the shadow and the worldline of something else (a piece of ground or wall, for example) that occurs at the moment when the car is scratched. In (b) there is an interaction between the child's arm and a beam of sunlight and this interaction transmits properties to the intervening air and ultimately to those molecules that interact with the shadow. Here there is little strain in talking about an *indirect* interaction between the child and the shadow. Thus, if the proposed remedy is to succeed, it will be necessary to explain why the *intersection* of spacetime paths that occurs in (a) does not count as a proper marking and why the apparent *indirect interaction* of (b) should be considered pathological. Notice that even if these explanations are given, there will still be difficulty in accounting for (d), for in that case there is a process (the projectile) that interacts with the shadow.

If (CP) is to be salvaged, then it will be a consequence of the fact that the constraints on causal interactions debar the kinds of pseudomarkings and derivative

markings that figure in (a) and (b). So let us investigate Salmon's explicit characterization of causal interactions. This is presented in a principle (CI) (1984, 171):

> (CI) Let P_1 and P_2 be two processes that intersect with one another at the spacetime point S, which belongs to the histories of both. Let Q be a characteristic that process P_1 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_1) if the intersection with P_2 did not occur; let R be a characteristic that process P_2 would exhibit throughout an interval (which includes subintervals on both sides of S in the history of P_2) if the intersection with P_1 did not occur. Then the intersection of P_1 and P_2 at S constitutes a causal interaction if:
>
> (1) P_1 exhibits the characteristic Q before S, but it exhibits a modified characteristic Q' throughout an interval immediately following S; and
>
> (2) P_2 exhibits the characteristic R before S, but it exhibits a modified characteristic R' throughout an interval immediately following S.

(CI) must be read as claiming only that there are some characteristics Q and R for which these conditions are met, since it would obviously be hopeless to propose that a causal interaction modify all the characteristics that a process would otherwise have retained. We should also note that it is irrelevant for present purposes to demand that the processes P_1 and P_2 be genuine causal processes. Presumably, if they are pseudoprocesses (or worse), they should not be able to satisfy these conditions in any case. However, since the distinction between causal processes and other spatio-temporal lines is what we are trying to draw, we could not appeal to the notion of causal process in (CI), even if we wanted to.

As we saw in the discussion of example (a) above, although when the car scrapes the wall there is no obvious process that can be singled out as intersecting the shadow, there are innumerable contrived spacetime paths that intersect the shadow at the same time. I sketched a strategy for resurrecting the counterexample: select one of these that changes some contrived characteristic just as the shadow takes on the property of being the shadow of a marked car. I shall now develop the strategy in a concrete way and show that (CI) fails to block it.

Assume, for the sake of simplicity, that the shadow of the car is being cast on the wall against which it scrapes, and that there is a stone in the wall, P_2, that is just touched by the shadow of the car at the moment when the car scrapes the wall. If the car had not been so close as to scrape the wall, then the shadow would not have touched that stone. Given that it scraped, then the stone had to be touched by the shadow. Let P_1 be the shadow. Then:

> P_1 and P_2 intersect at the moment at which the car scrapes the wall.
>
> Prior to the moment of intersection, P_1 has the characteristic of being the shadow of an unscratched car; after that moment, P_1 has the characteristic of being the shadow of a scratched car.

> Prior to the moment of intersection, P_2 has the characteristic of being untouched by the shadow; after the intersection, it has the characteristic of having been touched by the shadow.
>
> Had the intersection not occurred, then, by hypothesis, the car would not have scraped the wall, so that the shadow would not have taken on the property of being the shadow of a scratched car. Had the intersection not occurred, then the stone would not have taken on the characteristic of having been touched by the shadow.

Conclusion: the intersection between P_1 and P_2 meets all the conditions laid down in (CI) for a causal interaction. Hence, we cannot use (CI), Salmon's official explication of the concept of causal interaction, to dismiss the property of being the shadow of a scratched car as a pseudomark.

The discussion so far teaches us how to generalize the problem of which (a) is the prototype. If (CI) exhausts the conditions on causal interactions, then almost any intersection of continuous spacetime paths will constitute a causal interaction. For let P_1 and P_2 be any two continuous spacetime paths, both meeting the condition that for any pair of its points one belongs to the future light-cone of the other, which intersect just once at S. Apparently the intersection at S will count as a causal interaction. For, let Q be the characteristic of not having intersected P_2, R be the characteristic of not having intersected P_1, Q' be the characteristic of having intersected P_2, and R' be the characteristic of having intersected P_1. Then it appears that all the conditions of (CI) are satisfied and that the intersection is a *bona fide* causal interaction.

The most promising rejoinder to the line of argument that I have been developing seems to be to lean heavily on the appeal to counterfactuals. Can we make sense of the idea that, if the intersection with P_2 had not occurred, P_1 would have continued to have Q? That depends on how the processes are picked out. If the identity of P_1 consists in its being that very set of spacetime points (and similarly for P_2) then it is difficult to make sense of the antecedent of the counterfactual. We must impose one further condition to maintain the counterexample: the processes must be identifiable apart from the spacetime points that constitute them. In the case of pseudoprocesses, it is not hard to meet this extra requirement. Imagine, for example, that P_1 and P_2 are intersecting shadows.

Another way to try to save Salmon's account is to place some limitations on the kinds of properties that can serve as marks. Since Goodman's profound investigation of problems about counterfactuals and inductive projection, it has been evident that there are serious difficulties in circumscribing the *genuine* properties (or the genuine natural kinds). The literature on events and "Cambridge changes" (see Geach 1969, Kim 1974) exposes further troubles that must be addressed if cases like (a) and (b) are to be excluded by distinguishing gerrymandered properties and ruling them out as potential marks. Recognizing the kinship between ex-

amples like (a) and Goodman's deep puzzles may prove helpful in suggesting an alternative approach to the epistemology of causation, one that would attempt to emulate Goodman's own preferred line of solution to his "new riddle." I shall explore this possibility below.

At best, the maneuvers I have been considering will avail with (a) and its generalization. As already noted, (b) seems to involve some kind of interaction that produces a modification of the shadow, and (d) is an especially pure case in that the modification is direct. Here there seems to be no crucial dependence on contrived "causal interactions," on gerrymandered "marking properties," or on other kinds of Goodmanesque trickery. The shadow takes on a genuine new property: its shape is distorted. Moreover, the distortion would persist without any further interaction—in whatever sense we can give to the idea of macroscopic objects retaining their properties in the absence of further interactions, see (c)—because the projectile retains its position relative to the shadow without any application of force. The example exploits Salmon's strategy of linking causation to the world-picture provided by physics. We don't need further interactions when we have inertial motion, so we can keep a pseudoprocess marked without further interactions by arranging for the mark to undergo inertial motion with the pseudoprocess.

Is it really correct to insist that the mark (the distortion of shadow shape) is preserved without further interactions? After all, the continuation of the shadow requires a surface on which the shadow can be cast, and if this surface were cunningly contoured, there might be compensation for the distortion that the projectile would have produced in the shape of the shadow. But this is simply a special case of a point noted in connection with (c). Ordinary marks on ordinary causal processes—smudges on the cheeks of small children, pieces of chewing gum on balls—persist only if the interactions necessary for the persistence of the marked objects occur. Given sufficiently odd background conditions—driving rain on the child or lubrication of the surface of the ball—the mark would no longer persist. There seems to me to be no obvious basis for distinguishing the perfectly ordinary conditions that help to maintain the distortion of shadow shape from the perfectly ordinary conditions that help to maintain the chewing gum on the baseball or the smudge on the cheek.

Nor is it possible to protest that the interaction between projectile and shadow is a peculiar one in that the worldlines fuse after the intersection. Not only does Salmon explicitly allow for lambda-type interactions (cases in which processes fuse) but we could easily amend the example to ensure that part of the entering process should have a subsequent part of its worldline independent of the process that it intersects. We simply have to suppose that there is a part of the projectile that lands on the interior of the shadow and that is broken off and deflected by the collision. Of course, it would be unreasonable to demand that the *entire* entering process should emerge from the interaction, for that would be at odds with

the paradigm cases of marking that Salmon hopes to vindicate: when the chalk marks the object only part of the chalk persists after the interaction.

Cases like (d) provide another clue to the problem of the epistemology of causation. One very obvious point about the situation that I have envisaged is that all kinds of circumstances have to be just right. Consider the general class of cases in which projectiles land on the shadows of moving objects and distort them. In the vast majority of such episodes, the distortion is temporary. The projectile does not move with the shadow in such a way as to maintain the modification of shape. However, even if the velocities of car and projectile are not so exactly coordinated that the projectile maintains a constant relative position to the shadow, there may still be an interval of time throughout which the projectile slides across the shadow in such a way as to distort it. The fact that, at the end of the interval, the shadow reverts to its former shape should not be disconcerting. Many of our paradigms of marks are removed after a finite time in the normal course of events. But isn't it true that an uncoordinated projectile will produce *different* distortions at different points of the time interval during which it distorts the shadow? Yes, but equally, in the ordinary course of events, the shape and thickness of a chalk mark will alter (albeit in ways that are typically imperceptible to us) through the interval in which it marks a ball.

Yet our obvious point suggests something I take to be important. Suppose that we deployed a strategy of attempting to explain the characteristics of shadows by appealing to the properties of processes with which they intersect. That strategy would be doomed to failure because, in the typical case, the processes are not sufficiently well coordinated with the behavior of the shadow for us to be able to account for the subsequent features of the shadow in these terms. The best unification of our beliefs is obtained by incorporating within the explanatory store a pattern of argument in which we derive conclusions about shadow shape from premises about the motions of the objects with respect to sources of light that they partially block and about the characteristics of the surfaces on which the shadows are cast.[28] In general, the motions of objects that cross the shadow are relevant to the shape of the shadow only insofar as they affect the contours of the surface. We could not carry through a strategy of tracing changes in shadow shape to interactions with physical objects that cross the shadow – or, more precisely, we could not do this in a general way. Thus, while *local* violations of Salmon's principle are possible, the recognition that explanation is a matter of instantiating patterns that are applied *globally* enables us to diagnose the source of the trouble. I shall expand on this idea in subsequent sections.

So far I have proceeded by taking some common-sense ideas about macroscopic processes for granted. Strictly speaking, however, macroscopic processes do not conform to Salmon's conditions. The reason is that, as (c) suggests, the characteristics of a macroscopic object are (almost always) sustained by further interactions. Furthermore, the characteristics of such processes are often mod-

ified by subsequent interactions, so that, when a process is "marked" (in the ordinary sense) it is likely that the mark will be altered by a later interaction to produce a later mark, which is, in turn, altered to produce a later mark, and so on. Embryological marking furnishes a host of dramatic cases. The embryologist often injects into one of the early cells of an embryo some substance that reacts with the molecules of the cell to produce a new state in which batteries of genes may be switched on or off at quite different times. Consequently, concentrations of proteins may be significantly different from the way they would otherwise have been, so that there may be a new cellular state, an altered tissue geometry, modified intercellular interactions, altered states of neighboring cells, and so on in a cascade of anomalous effects until the biologist finally sees a deviant morphology.

We have here a sequence of causal processes and interactions, and, if we say that the final organism is marked by the initial interaction, that is because we envisage a sequence of marks such that each is transmitted by a causal process that interacts with another process to produce a successor mark. Our attribution is based on our acceptance of a chain of counterfactuals: if P_n had not transmitted M_n then P_{n+1} would not have acquired M_{n+1}. In building up a complex causal process out of elementary causal processes—that is, processes that do not interact with other processes—we need to make heavy use of counterfactuals. The sequence $P_1, \ldots, P_n$ constitutes a complex causal process only if each P_r interacts with P_{r+1} so that if P_r had not been modified to bear Q_r then P_{r+1} would not have been modified to bear Q_{r+1}.

Let us take stock. I have elaborated three kinds of difficulties for Salmon's project, some of which point to quite general troubles in providing an empiricist account of the justification of causal claims. First, there are problems in distinguishing the genuine causal processes from other continuous spatio-temporal paths, and problems in distinguishing causal interactions from mere intersections, all of which are based on the existence of unwanted (spurious?) characteristics—pseudomarks, pseudomodifications, and the like. Second, there are independent problems with both distinctions resulting from the possibility of exploiting the inertia of some processes that can be fortuitously coordinated with pseudoprocesses to "mark" them. Third, Salmon's conditions seem applicable only to ideal (elementary) processes. This requires that we provide an account of the way in which the causal structure of the macroscopic world results from the stringing together of elementary processes. Even if we already had such an account, the emerging picture of our causal knowledge is one in which the justification of *recherché* theoretical claims about idealized processes seems to be fundamental and our ordinary causal knowledge derivative. This is, of course, grist for a mill that I ground above (4.5), when I insisted that our everyday causal knowledge is based on our early absorption of the theoretical picture of the world bequeathed to us by our scientific tradition. But, in any case, as I have argued by reference

to the embryological example, our concept of an ordinary macroscopic process involves us in commitment to a large number of counterfactuals.

6.3 Causation and Counterfactuals

It is now time to take up what I view as the most serious trouble of Salmon's project and which I take to threaten any program that tries to use causal concepts to ground the notion of explanation while remaining faithful to an empiricist theory of knowledge. Salmon is very clear in acknowledging that he needs to appeal to counterfactuals in stating his principles (see, for example, 148–49, 171ff.) and in recognizing that he must show how the relevant counterfactuals can be justified. I shall try to show first that the counterfactual commitments of Salmon's theory are far more extensive than he has noted (indeed, I shall claim that the theory naturally evolves into a counterfactual theory of causation), and second that the epistemology for counterfactuals that Salmon outlines encounters grave difficulties.

Return to the question with which we began. We aim eventually to specify the conditions under which two events are causally related, and we have been assuming so far that this can be done by claiming that there is a causal process that links the two events with causal interactions occurring at either end. Now if this assumption were true, then causal explanation would be extraordinarily easy. Take any object that participates in the event for which we want to give an explanation. Its worldline will be a causal process. The considerations that we have drawn from (c) lead us to believe that there will be innumerable causal interactions along this worldline as we trace backward into the past. Any of these earlier points is a terminus of a causal process that is linked to the event that we hope to explain, and at both it and the *explanandum* event we have causal interactions.

The problem of finding causal structure is surely more difficult than this, and the difficulty stems from the fact that we must delineate the *right* causal processes and the *right* causal interactions. Our paradigm, the baseball that hits the window, turns out to be far more intricate than we might have thought. Suppose P is the process that constitutes the window, P_t the point at which the window shatters and P_t' some earlier point at which its temperature is changed through the impact of a sudden gust of wind. There is a causal process (the segment of P bounded by P_t' and P_t) that links P_t' and P_t. There are interactions at both ends: the interaction with the gust of wind at P_t' and the interaction with the ball at P_t. (There may even be other interactions at P_t—a second gust of wind, for example.) But the existence of this particular structure of interactions and process is not crucial to the causal history of the event of shattering. To specify the conditions under which c and e are causally related, we need to build into the account the idea that the initial interaction produces the modification that is responsible for the characteristics of the terminal interaction.[29] Intuitively, what is lacking is the kind of

articulated structure that I envisaged in the building up of complex processes out of simple ones.

Surely what we want to say about the baseball is something like this. There is an event, the spacetime point c, at which two processes (the bat and the ball) intersect (or, more exactly, overlap). As a result of the intersection, the characteristics of both processes are modified from what they would otherwise have been. Focus on the ball (P_1). We can explicate the dangerously causal-sounding phrase "as a result of" by offering the counterfactual

(A) If the bat had not intersected P_1 then the momentum of P_1 would have been different.

It is in virtue of (A) and related counterfactuals for both bat and ball that we count the intersection of bat and ball as a causal interaction.

Next consider the motion of the ball to the window. During its flight the ball engages in numerous interactions that modify its momentum. What occurs is very like the production of derivative marks in the case of the embryo. We believe many counterfactuals about all these interactions, but, for the sake of simplicity, they may be condensed into

(B) If the momentum of P_1 after its intersection with the bat had been different then the momentum of P_1 just prior to its intersection with P_2 (the window) would have been different.

Now it is because the ball has the momentum it does when it intersects the window that the window breaks in the way that it does. In other words

(C) If the momentum of P_1 just prior to its intersection with P_2 had been different then the properties of P_2 just after the intersection would have been different (specifically, the window would not have broken!).

I take it as evident that these particular counterfactuals can be replaced with weaker counterfactuals about the ball's having non-zero (or appreciable) momentum at different stages of its career. The crucial point is that our claim of a causal relation between c and e depends not simply on the existence of the interactions and the processes but on our acceptance of the counterfactuals (A)–(C) (or of related weaker versions). *We have to invoke counterfactual notions not only in characterizing the concepts of causal process and causal interaction but also in singling out the causal processes and causal interactions that are relevant to particular events.*

But if so much counterfactual machinery is needed to articulate the conditions under which events are causally related, do we really need to talk about processes and interactions at all? Suppose that we simply believed counterfactuals like (A)–(C)—or just that if the bat had not struck the ball then the window would not have been broken. Would this be enough to countenance a causal relation between

the striking of the ball and the breaking of the window? Would it be sufficient even if we doubted that there were any intervening process?

Entertaining this possibility takes us in the direction either of past science or of science fiction. The past science comes in with the idea of instantaneous action-at-a-distance. Some philosophers may hold that we would properly maintain that there is a causal relation between two widely separated events if we believed that (i) A acquired property P at t, (ii) B acquired property Q at t, and (iii) if A had not acquired P then B would not have acquired Q. But it is not easy to provide compelling reasons in support of this way of talking, as the eighteenth-century Newtonians discovered.

Science fiction is more promising. Time-travel can be defended as logically possible (see Lewis 1976, Horwich 1987), and certainly time-travel into the future seems less problematic than time-travel into the past. Imagine a time-traveler whose worldline terminates at t, just after he has eaten a peanut butter and jelly sandwich. The time-traveler reappears at a much later time t' with the peanut butter and jelly sandwich in his stomach. He has the sandwich in his stomach at t' because he was ingesting it just before t. There is no continuous process that links the two parts of his worldline, but the pertinent counterfactual holds: if he had not eaten the sandwich then it would not be in his stomach.

In this example there is no continuous process but there are two continuous fragments that surround the big gap in the time-traveler's life. We can make the situation even more dramatic by focusing on complex similarities that hold at the point of departure and at the point of return. As he departs, the time-traveler is quoting verse: "I warmed my hands before the fire of life/It sinks and I . . . "; as he reappears he is still quoting " . . . I am ready to depart." His lips are spaced as they are at the moment of return because that is the way they were when he left. If he had not been in the process of saying "I" when he left, he would not have been in the process of saying "I" when he returned.

I suggest that we can have causation without linking causal processes, and hence causal relations among events at which very peculiar interactions occur. What is critical to the causal claims seems to be the truth of the counterfactuals, not the existence of the processes and the interactions. If this is correct then it is not just that Salmon's account of the causal structure of the world needs supplementing through the introduction of more counterfactuals. The counterfactuals are the heart of the theory, while the claims about the existence of processes and interactions are, in principle, dispensable. Perhaps these notions may prove useful in protecting a basically counterfactual theory of causation against certain familiar forms of difficulty (problems of pre-emption, overdetermination, epiphenomena, and so forth).[30] But, instead of viewing Salmon's account as based on his explications of process and interaction, it might be more revealing to see him as developing a particular kind of counterfactual theory of causation, one that has some extra machinery for avoiding the usual difficulties that beset such proposals.

6.4 Justifying Counterfactuals

But of course many empiricists, including Salmon, worry about the use of counterfactuals in providing accounts of causation. Their anxieties need not simply be the result of an over-reverential attitude toward Hume, Carnap, and their successors. As Goodman (1956) made clear, the straightforward ways of providing a semantics for counterfactuals that will provide them with truth conditions whose presence or absence can be ascertained in ways that empiricists take to be unproblematic encounter apparently insuperable obstacles. Of course, there are illuminating treatments of the semantics of counterfactuals (for example, Lewis 1974, Stalnaker 1968) that enable us to appreciate many points about the logic of counterfactuals that had eluded earlier workers and to formulate controversial issues that had previously been missed. But one thing that semantical accounts in terms of possible worlds and similarity relations fail to do for us is to provide truth conditions that are epistemologically unproblematic. Indeed, the situation with respect to counterfactuals (and modality generally) is distressingly similar to a predicament in the philosophy of mathematics that Benacerraf (1973) presents in compelling fashion: our best semantic accounts and our best epistemological views do not cohere.[31] For the best semantic accounts make reference to possible worlds, our best epistemological views make knowledge (and justification) dependent on the presence of natural processes that reliably regulate belief, and it is (to say the least) unobvious how any natural process could reliably regulate our beliefs about possible worlds.

Salmon does not try to give a semantics for counterfactuals or to resolve the dilemma that I have just indicated. Nor does he endeavor to explain how the counterfactual conditional is reducible to some privileged (empiricistically acceptable) notions. Instead, he pursues the more modest task—which is sufficient for his purposes—of offering an account of how we might justify beliefs in counterfactuals. His account has the virtue of providing a convincing description of the ways in which people do in fact try to justify counterfactuals: no esoteric processes or faculties are invoked. Salmon suggests that we base our counterfactual knowledge on the method of control experiments. To test "If A had not been, then B would not have been" we are to take a sample of test entities, divide it into two subsets under similar conditions, subject one subset to the presence of A and the other to the absence of A. But, of course, the trouble with counterfactuals is that you cannot hold fixed *all* the circumstances of the antecedent while making the antecedent (which is actually false) come true. You have to be selective about what is held constant, and the selection may make a difference to the outcome of the test. Even if the control group and the test group are "similar," the fact that they differ with respect to the presence or absence of A will mean that they differ with respect to many other characteristics. How do we design the *valid* tests,

those that do not introduce some crucial difference, unrelated to the presence or absence of A, that gives a misleading result?

We can appreciate the general problem by considering a concrete case, arising from an example that Salmon considers (the example is originally due to Patrick Maher and Richard Otte). Imagine that two billiard balls roll across a table with a transparent surface, and that they collide. The table is illuminated from above, so that there is also an intersection of their shadows. Under these circumstances, we will want to claim that the impact of Ball A on Ball B causes a change in the direction of motion of Ball B, but that the crossing of the shadow of Ball B by the shadow of Ball A does not cause the change in the direction of motion of the shadow of Ball B. The problem is to explain how the first causal claim is justified and the second is not.

Call the shadow of Ball A P_A and the shadow of Ball B P_B. Then the problem arises because, at first sight, both of the following counterfactuals are true

(i) If the worldline of Ball A had not intersected that of Ball B, then the direction of motion of Ball B would not have altered.

(ii) If the worldline of P_A had not intersected that of P_B, then the direction of motion of P_B would not have altered.

Following the method of testing that Salmon suggests, we might confirm (i) by dividing a sample of 100 ball rollings into two types—cases in which there is an intersection of balls on the table (the control cases) and cases in which there is no such intersection (the experimental cases). Call this Experiment I. Allegedly Experiment I confirms (i) because in the cases in which the balls do not meet their directions are unaltered. Now when the balls do not meet neither do their shadows. Hence we might suppose that Experiment I also suffices to confirm (ii)—for does it not show that when the shadows do not meet the directions of motion of the shadows are unaltered?

Salmon claims that Experiment I is the wrong experiment for testing (ii). He urges that our experimental group should comprise cases in which the balls collide, but part of the surface of the floor is illuminated from below the table so that P_A is absent. Call this Experiment II. In Experiment II P_B changes its direction of motion, even in the experimental cases, so we have grounds for saying that II refutes the crucial counterfactual.

The important question is how we justify testing (ii) by performing Experiment II rather than Experiment I. Salmon is surely right to claim that II is the intuitively right experiment to perform, the experiment that accords with our natural ideas about the causal character of the situation. But we are looking for the basis of those ideas, and, in consequence, we need a theory that tells us which controlled experiments are the correct ones.

To see the force of this question, consider an analogous experiment that is proposed to test (i). In the experimental cases, we remove Ball A from the table but

we "simulate its presence" by subjecting Ball *B* to an appropriately strong impulse at the moment when it would have collided with the absent ball. (Note that the impulse restores certain similarities with the control cases that would otherwise be absent: for example in the motions of air molecules above the table.) Call this Experiment III. Now it is natural to protest that Experiment III involves gratuitous intervention. But isn't it possible to make a similar protest in the case of Experiment II where we introduce a new object (the light source) into the situation? What justifies us in thinking that the interventions involved in II do not confound the probative force of the test while those involved in III do?

The obvious answer (very obvious to anyone who thinks about counterfactuals along Lewis-Stalnaker lines) is that we try to keep the situation of the control group as close to that of the experimental group as possible. But in which respects should we prize similarity? In Experiment I the test group and the control group involve two balls and two shadows; the difference is just that, in the cases in the control group, the balls collide while in those of the test group they do not. In Experiment II the features of the balls, including their collision, are common to both groups; but the groups differ in that the test group shows only one shadow and involves an extra object, the new light source. Finally, in the crazy Experiment III we also introduce a new entity, the source of the impulse, but we may restore certain kinds of similarity between the control and the experimental cases (for example in the air currents above the table). There is a tradeoff here of similarities and differences, and we do not have a theory of the justification of counterfactuals until we have an account of how the trades are made.

In practice, of course, scientists design control experiments by drawing on their background causal knowledge. They endeavor to ensure that the control group and the experimental group are similar in those respects *that they take to be potentially causally relevant*. Once we have some causal knowledge (perhaps a significant amount) then that causal knowledge can be used in the design of control experiments that will test counterfactuals in just the way that Salmon proposes.[32] But if we are looking for a theory of how we justify counterfactuals from scratch, then the appeal to the method of controlled experiments is of no avail.

6.5 Changing the Epistemological Framework

At this point it is reasonable to scrutinize the general epistemological assumptions which have been taken for granted in the framing of our problem, to see if the task that has been set is overly ambitious. For Hume and his logical empiricist successors, providing an account of our causal knowledge required identifying judgments that can be justified on the basis of observation and inferences that would lead from those judgments to justify the causal claims that we accept. However, there are reasons, forcibly presented by Quine, Sellars, Kuhn, and others, for rejecting the tacit idea that each person's knowledge can be reconstructed to

reveal a chain of justifying inferences whose ultimate premises are statements that are completely justified by the person's own experiences. Instead, each of us inherits a body of lore from the previous generation, and some, the creative and the talented, modify that corpus and thus affect what is transmitted to the successor generation. There is no question of justifying our knowledge from privileged premises that record our own observations, for each of us is thoroughly indebted to our contemporaries and to the historical tradition.

If we apply this general point to the issue of how we are able to have justifications for causal and counterfactual claims, then we should want to question the idea that somehow each of us begins without any causal knowledge (or any knowledge of counterfactuals) and manages to build up such knowledge without assistance from others. Instead, we are never in the predicament that Hume and the logical empiricists depict. From the start of our conscious experience we absorb causal judgments as part of the lore of our ancestors. Hence it is entirely reasonable to attack the problem of causal knowledge by describing a method for extending such knowledge, *provided that we can show how that method could have been used to build up the knowledge we now take ourselves to possess from the basis with which our prehistoric ancestors began.*[33]

The view I am recommending is exactly parallel to the account of mathematical knowledge that I have defended elsewhere (Kitcher 1983, chaps. 6–10). If we are to understand how the causal and counterfactual claims that we currently make are warranted then we should show that there is a sequence of states of science, beginning with some state for which we can trace a direct justification and ending with our own corpus, such that each state in the sequence is obtained from its predecessor by a justification-conferring transition. Now it is possible that defenders of the causal approach to explanation can complete the project I have outlined, showing how causal knowledge is systematically built up without appealing to methodological principles that govern the acceptance and rejection of putative explanations. The aim of this section has been to identify the problems that they will have to overcome, not to close the books on the causal approach. But it does seem to me that the general epistemological approach that I have been sketching is far more congenial to the view that explanation consists in the systematization of our beliefs.

Consider the following story, highly fanciful and oversimplified, but, I think, one that bears an important moral about the growth of our causal knowledge. Imagine our remote ancestors with a primitive stock of unconnected beliefs about the world. In attempting to systematize their beliefs they arrive at the first picture of the order of nature, coming to see some phenomena as dependent on others. At this stage, their state can be represented by the language they use, the claims that they accept, and the explanatory patterns that they endorse. Successive states of science are generated as the language is modified, the body of claims revised, and the store of explanatory patterns altered. From the very beginning the con-

struction of the explanatory store is guided by the directive to unify belief, in the sense given in section 4 (see sections 7 and 8 for further elaboration of it). At each stage, the explanatory store supplies an ordering for the phenomena and serves as the basis for the introduction of causal concepts. These are absorbed in childhood, sometimes with the giving of scraps of causal information, sometimes simply by learning parts of the language. But the crucial point is that the "because" of causation is always derivative from the "because" of explanation. In learning to talk about causes or counterfactuals we are absorbing earlier generations' views of the structure of nature, where those views arise from their attempts to achieve a unified account of the phenomena.

The final sections will attempt to defend one central idea of this story. I shall try to show that the principle of unification can be formulated so as to give genuine methodological guidance, enabling us to justify modifying the explanatory store in certain ways and not in others. *En passant*, I shall consider the problems of asymmetry and irrelevance.

7. Comparative Unification

There are two distinct contexts in which a methodological principle directing us to unify our beliefs can be expected to operate. In one of these—the simpler of the two—we consider a fixed body of beliefs and use the principle to select that set of derivations (among some class of alternative sets of derivations) that best unifies the belief corpus. The second context is more difficult because it allows for changes in the corpus, and possibly even in the language in which the beliefs are framed. However, we need to consider this context because changes in the state of a science, including changes in belief and in language, are often justified by appeal to the idea that the changes will yield an increase in explanatory power. I shall start with the simpler case, and introduce complications later.

7.1 Comparative Unification without Change of Belief

Let K be the set of statements accepted at some stage in the development of science, and let L be the language used to formulate those statements. Suppose that S and S' are sets of derivations such that all members of each set are acceptable relative to K. The principle that we want to adopt is

(U) S should be chosen over S' as the explanatory store over K, $E(K)$, just in case S has greater unifying power with respect to K than S'.

Note that (U) is formulated as a comparative principle, directing us to make a choice between proposed alternatives. This is because the project is to understand how justified choices are made in the growth of scientific knowledge, and we can safely assume that each choice situation involves a set of proposed alternatives.

The next task is to say how we make judgments of unifying power. Recall from

section 4.3 that unifying power depends on paucity of patterns used, size of conclusion set, and stringency of patterns. As I admitted, when the criteria pull in different directions, it will be hard to see how tradeoffs are made. However, this possibility *may* not prove troublesome. Perhaps whenever there is competition between two sets of patterns with different virtues, we can find an acceptable way to combine the virtues. Let us formulate an explicit principle to express optimism about this.

(O) Let U, U' be sets of patterns. Then there is a set of patterns U^* such that

(a) there is a one-one mapping from U^* to U, f, and a one-one mapping from U^* to U, f', such that for each pattern p in U^*, p is at least as stringent as $f(p)$ and at least as stringent as $f'(p)$; (one or both of f, f' may be injections rather than surjections)

(b) let S, S', S^* be the sets of derivations that are the complete instantiations of U, U', and U^* with respect to K; then the consequence sets $C(S)$, $C(S')$, $C(S^*)$ are such that $C(S)$ and $C(S')$ are both subsets (not necessarily proper) of $C(S^*)$.

(A bijection is a one-one mapping. A bijection from A to a subset of B is an injection into B. A bijection that is onto B is a surjection.)

Here clause (a) tells us that U^* does at least as well as its rivals by the criteria of stringency and paucity of patterns and (b) tells us that it does at least as well as generating consequences over K. If (O) is correct—or if it is correct for the bodies of belief that we are interested in considering, then the problem of tradeoffs is unworrying, because in a situation in which rival systematizations have different virtues we can always reject both of them in favor of a systematization that combines their merits. However, as my dubbing of the principle hints, I do not know whether (O)—or some useful restriction of it—is true.

We can formulate a condition on the comparative unifying power of sets of patterns in the obvious way.

(C) Let U, U' be sets of patterns and S, S' their complete instantiations with respect to K. Then U has greater unifying power than U' if one (or both) of the following conditions is met.

(C1) $C(S')$ is a subset of $C(S)$, possibly though not necessarily proper, and there is a one-one mapping f from S to S' such that for each pattern p in S, p is at least as stringent as $f(p)$, and such that either f is an injection or f is a surjection and there is at least one pattern p in S such that p is more stringent than $f(p)$.

(C2) $C(S')$ is a proper subset of $C(S)$ and there is a one-one map f from S to S' (either an injection or a surjection) such that for each p in S, p is at least as stringent as $f(p)$.

(C1) applies if S uses fewer or more stringent patterns to generate the same conclusions as S'. (C2) holds if S does equally well as S' by criteria of stringency and paucity of patterns and is able to generate a broader class of consequences. It is not hard to show that the comparative relation introduced by (C) has the right features to order sets of patterns with respect to unifying power. It is both asymmetric and transitive.

So far I have taken the notion of stringency for granted. As with my approach to comparative unifying power, I shall ignore the problem of tradeoffs and offer conditions for comparing argument patterns with the same structure. Consider first patterns that have a common classification. In this case, one is more stringent than another if corresponding schemata in the first are subject to demands on instantiation that are more rigorous than those in the second. The idea can be made more precise as follows.

(T) Let $<s,i>$ be a pair whose first member is a schematic sentence and whose second member is a complete filling instruction for that sentence, and let $<s',i'>$ be another such pair. Suppose that s and s' have a common logical form. Let g be the mapping that takes each nonlogical expression (or schematic letter) in s to the nonlogical expression (or schematic letter) in the corresponding place in s'. For any schematic letter t occurring in s, $<s,i>$ is tighter than $<s',i'>$ with respect to t just in case the set of substitution instances that i allows for t is a proper subset of the set of substitution instances that i' allows for $g(t)$; $<s,i>$ is at least as tight as $<s',i'>$ with respect to t just in case the set of substitution instances that i allows for t is a subset of the set of substitution instances that i' allows for $g(t)$. $<s,i>$ is tighter than $<s',i'>$ just in case, (i) for every schematic letter occurring in s, $<s,i>$ is at least as tight as $<s',i'>$ with respect to that schematic letter, (ii) there is at least one schematic letter occurring in s with respect to which $<s,i>$ is tighter than $<s',i'>$ or there is a nonlogical expression e occurring in s such that $g(e)$ is a schematic letter, and (iii) for every schematic letter t occurring in s, $g(t)$ is a schematic letter. If only conditions (i) and (iii) are satisfied, then $<s,i>$ is at least as tight as $<s',i'>$

Let p, p' be general argument patterns sharing the same classification. Let $<p_1, \ldots p_n>$ and $<p'_1, \ldots p'_n>$ be the sequence of schematic sentences and filling instructions belonging to p and p' respectively. Then p is more stringent than p' if for each j $(1 \leq j \leq n)$ p_j is at least as tight as p'_j and there is a k such that p_k is tighter than p'_k.

There is another way in which one argument pattern might be more stringent than another. One pattern might have a classification that indicated that an inferential transition is to be made by appealing to certain kinds of principles while another might articulate the intervening structure by specifying schematic premises that are to be linked in definite ways.[34] If the latter precludes certain

possible instantiations that the former leaves open, then it is appropriate to count it as more stringent. Once again, the idea can be made precise as follows.

(R) Let p, p' be general argument patterns such that the sequence of schematic sentences and filling instructions of p is $<p_1, \ldots p_n>$ and the sequence of schematic sentences and filling instructions of p' is $<p_1, \ldots p_r, q_1, \ldots q_s, p_{r+1}, \ldots p_n>$. Suppose that the classifications differ only in that for p one or more of the p_{r+j} is to be obtained from previous members of the sequence by derivations involving some further principles of a general kind G, while for p' that (or those) p_{r+j} are to be obtained from the same earlier members of the sequence and from some of the q_k by specified inferential transitions. Suppose further that in each case of difference the set of subderivations allowed by p' is a subset of the set of subderivations allowed by p, and that in at least one case the relation is that of proper inclusion. Then p' is more stringent than p.

(T) and (R) together provide analyses of the two basic ways in which one argument pattern may be more stringent than another. I do not pretend that they provide a complete account of relative stringency, but I hope it will be possible to see how to combine them. Of course, there is the worrying theoretical possibility that we may be forced to judge between argument patterns, one of which scores well by the kinds of considerations adduced in (T), while the other is recommended by the kinds of considerations adduced in (R). As with the account of comparative unification, we may hope that when this occurs there will be some acceptable argument pattern that combines the merits of both—but perhaps this is overly optimistic. In any case, the conditions I have given enable us to tackle the problems of comparison that we need to address.

At this point we have the resources to understand how the justification of accepting certain sets of derivations, rather than others, might work. Suppose that we are comparing the merits of two systematizations of K, S and S'. If the best set of patterns we can think of that will generate S fares better in unifying K, as judged by (C) in light of such claims about stringency as (T) and (R), than the best set of patterns we can think of that will generate S', then, according to (U), we are entitled to prefer S to S'. Note that we should allow for justifiable mistakes in overlooking some relatively recondite complete generating set, just as we should allow for justifiable error on the part of a scientist who misses some subtle idea in formulating a theoretical view.

7.2 The Possibility of Gerrymandering

Unfortunately, there is a serious worry about whether the principles that I have introduced will enable us to solve the problems of asymmetry and irrelevance by debarring unwanted derivations. Intuitively, the line of solution that we want to

adopt consists in showing that those who accept the wrong derivations are committed to accepting more patterns than they need, or to accepting less stringent patterns than they should, or to generating a more restricted set of consequences. But, as matters now stand, there is room to wonder whether, given any derivation that one wants to incorporate within the explanatory store, one can always discover a set of derivations including it and contrive a set of patterns that will serve as the basis of that set of derivations, in such a way that the contrived set of patterns will score just as well as any rival set of patterns that will generate the derivations that the orthodox accept. To put it bluntly, are the principles that I have assembled toothless?

The difficulty arises because the account I have given lacks resources for convicting those who employ deviant derivations of using deviant patterns. We can formulate a general strategy for exploiting the deficiency as follows. Let us suppose that our accepted ideal explanation of some *explanandum* E is a derivation d that instantiates a pattern p. Someone now proposes that a different derivation d' furnishes the ideal explanation of E. As we shall see shortly, the general strategy of challenging deviance is to show that acceptance of d' would commit us to employing a pattern p', confronting us with the options of either generating the more limited set of consequences yielded by p' or else employing both p and p' and thereby flouting the maxim of minimizing the number of patterns used. But this strategy depends crucially on a view about how patterns are to be individuated. Imagine that the set of derivations proposed as a rival to orthodoxy is just the orthodox set except that d' has been substituted for d. How do we show that the rival set has to use one more pattern than the orthodox set in generating the same consequences?

The obvious and natural approach is this. Call the orthodox set S, the rival S'. S contains many instantiations of p other than d, and S' will contain these derivations too. Hence, if p belongs to the basis of S then p must belong to the basis of S'. But d' doesn't instantiate p, so it must be generated from some other pattern p' in the basis of S'. Now p' cannot belong to the basis of S, for, if it did, then d' would belong to S (recall that an acceptable systematization must contain all instantiations of the patterns in its basis). Therefore p' must be an extra pattern, so that the basis of S' contains all the patterns in the basis of S and one more besides.

But this cannot be quite right. For if S' has *precisely* the pattern p in its basis, then, provided we heed the requirement of completeness, all instantiations of p, and hence d itself, would have to belong to S'. So the basis of S' must contain some doctored version of p, p^d, that yields all the instances of p except for d. Presumably this is accomplished by adding some filling instruction to debar the substitutions that would generate d. Now, of course, it begins to look as though the basis of S' will lose on the paucity of patterns but win on stringency. But this is not the end of the matter. For while we are doctoring, we can surely contrive

some gerrymandered pattern $p^{d'}$ that will generate all the instantiations of p except for d together with d'. In effect, we would play with the classification and the schematic sentences until the derivations looked as alike as possible and then contrive filling instructions to let us substitute in just the right way to give the instantiations we want.

We need some requirements on pattern individuation that will enable us to block the gerrymandering of patterns by disjoining, conjoining, tacking on vacuous premises and so forth. The strategy sketched in the last paragraph attempts to disguise two patterns as one, and it does so by making distinctions that we take to be artificial and by ignoring similarities we take to be real. Thus the obvious way to meet the challenge is to demand that the predicates occurring in the schematic sentences, those employed in formulating the filling instructions, and those that figure in the classification all be projectable predicates of the language in which K is formulated. If a pattern fails to satisfy these conditions, then we must decompose it into several elementary patterns that do, and use the number of elementary patterns in our accounting.

7.3 Asymmetry and Irrelevance

The time has now come to put all this abstract machinery to work. Problems of asymmetry and irrelevance take the following general form. There are derivations employing premises which are (at least plausible candidates for) laws of nature and that fail to explain their conclusions. The task is to show that the unwanted derivations do not belong to the explanatory store over our current beliefs. To complete the task we need to argue that any systematization of our beliefs containing these derivations would have a basis that fares worse (according to the principles (U), (C), (T), and (R) stated above) than the basis of the systematization that we actually accept. In practice, this task will be accomplished by considering a small subset of the explanatory store, the derivations that explain conclusions akin to that of the unwanted derivation, and considering how we might replace this subset and include the unwanted derivation. I want to note explicitly that there is a risk that we shall overlook more radical modifications of the explanatory store which would incorporate the unwanted derivation. If there are such radical modifications that do as well by the criteria of unifying power as the systematization we actually accept, then my account is committed to claiming that we were wrong to treat the unwanted derivation as nonexplanatory.

7.3.1 The "Hexed" Salt

Let us start with the classic example of explanatory irrelevance. A magician waves his hands over some table salt, thereby "hexing" it. The salt is then thrown into water, where it promptly dissolves. We believe that it is not an acceptable

explanation of the dissolving of the salt to point out that the salt was hexed and that all hexed salt dissolves in water. What is the basis of this belief?

Suppose that *E(K)* is the explanatory store over our current beliefs, *K*, and that *S* is some set of derivations, acceptable with respect to *K*, that has the unwanted derivation of the last paragraph as a member. One of the patterns used to generate *E(K)* derives claims about the dissolving of salt in water from premises about the molecular composition of salt and water and about the forming and breaking of bonds. This pattern can be used to generate derivations whose conclusions describe the dissolving of hexed salt and the dissolving of unhexed salt. How does *S* provide similar derivations? Either the basis of *S* does not contain the standard pattern or it contains both the standard pattern and a nonstandard pattern that yields the unwanted derivation. In the former case, *S* fares less well than *E(K)* because it has a more restricted consequence set, and, in the latter case, it has inferior unifying power because its basis employs all the patterns of the basis of *E(K)* and one more besides.

It is obviously crucial to this argument that we exclude the gerrymandering of patterns. For otherwise the claim that the basis of *S* must contain either the nonstandard pattern alone or the nonstandard pattern plus the standard pattern would be suspect. The reason is that we could gerrymander a "pattern" by introducing some such Goodmanian predicate as "*x* is either hexed, or is unhexed and has molecular structure *NaCl*." Now we could recover derivations by starting from the claim that all table salt satisfies this predicate, by using the principle that all hexed table salt dissolves in water to generate the conclusion from one disjunct and by using the standard chemical derivation to generate the conclusion from the other disjunct. This maneuver is debarred by the requirement that the predicates used in patterns must be projectable from the perspective of *K*.

Consider next a refinement of the original example. Not all table salt is hexed, but presumably all of it is hexable. (For present purposes, we may assume that hexing requires only that an incantation be muttered with the magician's thoughts directed at the hexed object; this will obviate any concerns that some samples of table salt might be too large or too inaccessible to have the magician wave a hand over them.) Suppose now that it is proposed to explain why a given sample of table salt dissolves in water by offering the following derivation:

a is a hexable sample of table salt.
a was placed in water.
Whenever a hexable sample of table salt is placed in water, it dissolves.
a dissolved.

I take it that this derivation strikes us as nonexplanatory (although it is useful to point out that it is not as badly nonexplanatory as the derivation in the original example). Suppose that *S* is a systematization of *K* that contains the derivation. Can we show that *S* has less unifying power than *E(K)*?

Imagine that *S* had the same unifying power as *E(K)*. Now in *E(K)* the mini-derivation that is most akin to the one we want to exclude derives the conclusion that *a* dissolved from the premise that *a* is a sample of table salt, the premise that *a* was placed in water, and the generalization that samples of table salt that are placed in water dissolve. Of course, this mini-derivation is embedded within a much more exciting chemical derivation whose conclusion is the generalization that samples of table salt dissolve when placed in water. That derivation instantiates a general pattern that generates claims about the dissolving (or failure to dissolve) of a wide variety of substances from premises about molecular structure. In its turn, that general pattern is a specification of an even more general pattern that derives conclusions about chemical reactions and state changes for all kinds of substances from premises about molecular structures and energy distributions. If *S* is to rival *E(K)* then it must integrate the unwanted mini-derivation in analogous fashion.

That can be done. One way to proceed would be to use the standard chemical derivation to yield the conclusion that all samples of table salt dissolve when placed in water and then deduce that all hexable samples of table salt dissolve when placed in water. But now we can appeal to a principle of simplifying derivations to eliminate redundant premises or unnecessary steps. When embedded within the standard chemical derivation, the unwanted mini-derivation is inferior to its standard analog because the latter is obtainable more directly from the same premises. An alternative way of trying to save the unifying power of *S* would be to amend the standard chemical patterns to suppose that they apply only to hexable substances. But since it is supposed that *all* substances are hexable, and since this fact is used throughout *S* to generate derivations to rival those produced in *E(K)*, this option effectively generates a set of derivations that systematically contain idle clauses. Since it is believed that everything is hexable, the outcome is as if we added riders about objects being self-identical or being nameable to our explanations, and again a principle of simplification directs that the idle clauses be dropped.[35]

We can now achieve a diagnosis of the examples of explanatory irrelevance. Citation of irrelevant factors will either commit one to patterns of explanation that apply only to a restricted class of cases or the irrelevancies will be idle wheels that are found throughout the explanatory system. The initial hexing example illustrates the first possibility; the refinement shows the second.

7.3.2 Towers and Shadows

Let us now turn to the asymmetry problem, whose paradigm is the case of the tower and the shadow. Once again, let *K* be our current set of beliefs, and let us compare the unifying power of *E(K)* with that of some systematization *S* containing a derivation that runs from the premises about shadow length and sun eleva-

tion to a conclusion about the tower's height. As in the case of the irrelevance problem, there is a relatively simple argument for maintaining that *S* has less unifying power than *E(K)*. There are also some refinements of the original, troublesome story that attempt to evade this simple argument.

Within *E(K)* there are derivations that yield conclusions about the heights of towers, the widths of windows, the dimensions of artefacts and natural objects alike, which instantiate a general pattern of tracing the present dimensions to the conditions in which the object originated and the modifications that it has since undergone. Sometimes, as with flagpoles and towers, the derivations can be relatively simple: we start with premises about the intentions of a designer and reason to an intermediate conclusion about the dimensions of the object at the time of its origin; using further premises about the conditions that have prevailed between the origin and the present, we reason that the object has persisted virtually unaltered and thus reach a conclusion about its present dimensions. With respect to some natural objects, such as organisms, stars, and mountain ranges, the derivation is much more complex because the objects have careers in which their sizes are substantially affected. However, in all these cases, there is a very general pattern that can be instantiated to explain current size, and I shall call derivations generated by this pattern *origin-and-development* explanations.

Now if *S* includes *origin-and-development explanations*, then the basis of *S* will include the pattern that gives rise to these derivations. To generate the unwanted derivation in *S*, the basis of *S* must also contain another pattern that derives conclusions about dimensions from premises about the characteristics of shadows (the *shadow* pattern). In consequence, *S* would fare worse than *E(K)* according to our principles ((U) and so forth) because its basis would contain all the patterns in the basis of *E(K)* and one more. Notice that, once again, the "no gerrymandering" requirement comes into play to block the device of fusing some doctored version of the pattern that generates *origin-and-development* explanations with the shadow pattern. So *S* must foreswear *origin-and-development* explanations.

However, it now seems that *S* must have a consequence set that is more restricted than that of *E(K)*. The reason is that the *shadow* pattern cannot be instantiated in all the cases in which we provide *origin-and-development* explanations. Take any unilluminated object. It casts no shadow. Hence we cannot instantiate the *shadow* pattern to explain its dimensions.

This is correct as far as it goes, but the asymmetry problem cuts deeper. Suppose that a tower is actually unilluminated. Nonetheless, it is possible that it should have been illuminated, and if a light source of a specified kind had been present and if there had been a certain type of surface, then the tower would have cast a shadow of certain definite dimensions. So the tower has a complex dispositional property, the disposition to cast a shadow of such-and-such a length on such-and-such a surface if illuminated by a light-source at such-and-such an ele-

vation above the surface. From the attribution of this dispositional property and the laws of propagation of light we can derive a description of the dimensions of the tower. The derivation instantiates a pattern, call it the *dispositional-shadow* pattern, that is far more broadly applicable than the *shadow* pattern.

But can it be instantiated widely enough? To be sure it will provide surrogates for *origin-and-development* explanations in those cases in which we are concerned with ordinary middle-sized objects. But what about perfectly transparent objects (very thin pieces of glass, for example)? Well, they can be handled by amending the pattern slightly, supposing that such objects have a disposition to be coated with opaque material and then to cast a shadow. Objects that naturally emit light can be construed as having a disposition to have their own light blocked and then to cast a shadow. Objects that are so big that it is hard to find a surface on which their shadows could be cast (galaxies, for example) can be taken to have the dispositional property of casting a shadow on some hypothetical surface.

Yet more dispositional properties will be needed if we are to accommodate the full range of instances in which *origin-and-development* explanations are available. An embryologist might explain why the surface area of the primitive gut (archenteron) in an early embryo is of such-and-such a size by deriving a description of the gut from premises about how it is formed and how modified. To instantiate the *dispositional-shadow* pattern in such cases, we shall need to attribute to the gut-lining a dispositional property to be unrolled, illuminated, and thus to cast a shadow. A biochemist might explain the diameter in the double helix of a DNA molecule by identifying the constraints that the bonding pattern imposes on such molecules both as they are formed and as they persist.[36] Taking a clue from the principles of electronmicroscopy, the *dispositional-shadow* pattern can be instantiated by supposing that DNA molecules have a dispositional property to be coated and irradiated in specified ways and to produce absorption patterns on special surfaces. And so it goes.

Perhaps there are some objects that are too small, or too large, too light sensitive, or too energetic for us to attribute to them any disposition to cast anything like a shadow. If so, then even with the struggling and straining of the last paragraph, the *dispositional-shadow* pattern will still fail to generate derivations to rival those present in *E(K)*. But I shall assume that this is not so, and that for any object whose dimensions we can explain using our accepted patterns of derivation, it is possible to find a dispositional property that has something to do with casting a shadow.

However, if we now consider the critical predicate that appears in the *dispositional-shadow* pattern, we find that it is something like the following: "x has the disposition to cast a shadow if illuminated by a light source or x has the disposition to produce an absorption pattern if x is suitably coated and irradiated or x has the disposition to cast a shadow if x is covered with opaque material or x has the disposition to cast a shadow if x is sectioned and unrolled or x has the

disposition to cast a shadow after x has been treated to block its own light sources or . . . " At this point it is surely plain that we are cutting across the distinctions drawn by the projectable predicates of our language. Any "pattern" that employs a predicate of the sort that I have (partially) specified is guilty of gerrymandering, for, from our view of the properties of things, the dispositions that are lumped together in the predicate are not homogeneous. I conclude that even if it is granted that we can find for each object some dispositional property that will enable us to derive a specification of dimensions from the ascription of the disposition, there is no *common* dispositional property that we can employ for *all* objects. To emulate the scope of $E(K)$, the basis of S would have to contain a multiplicity of patterns, and our requirement against gerrymandering prohibits the fusion of these into a single genuine pattern.

As in the case of the irrelevance problem, there is a natural diagnosis of the trouble that brings out the central features of the foregoing arguments. Explanation proceeds by tracing the less fundamental properties of things to more fundamental features, and the criterion for distinguishing the less from the more fundamental is that appeal to the latter can be made on a broader scale. Thus an attempt to subvert the order of explanation shows up in the provision of an impoverished set of derivations (as in our original example of the tower and the shadow) or in the attempt to disguise an artificial congeries of properties as a single characteristic (as in our more recent reflections).

7.3.3 When Shadows Cross

At this point it is worth applying the account to the difficulties that emerged in the previous section. Consider the question why we do not consider the intersection of two shadows a causal interaction. The answer is surely that explaining the changes in directions of motion and in shape of a shadow by reference to the relations between it and another shadow would commit us to a pattern of explanation that is far less broadly applicable than that which consists in deriving the properties of shadows from the properties of the objects of which they are the shadows, in accordance with principles about light propagation. To see this, one need only note that we are able to instantiate the latter in situations where there is only one shadow around. Thus, a policy of explaining changes in the properties of shadows in terms of "interactions" among shadows would commit us to a pattern that would yield an impoverished explanatory store—unless, of course, we either admitted the standard pattern as well (thus adopting one more pattern than we need) or else gerrymandered a "pattern" by cutting across the divisions made within our language. I hope it will be apparent that many of the troubles that arose for Salmon's treatment of pseudoprocesses, pseudointeractions, and counterfactuals can be dissolved by invoking our principles about explanatory unification.

If the foregoing is correct, then the unification approach can apparently over-

come its most significant obstacle, namely the problems posed by explanatory asymmetry and explanatory irrelevance. But we should not celebrate too soon. As I remarked at the beginning of this section, the principle of explanatory unification should be expected to operate in two distinct contexts. We have been operating within the simpler of these, supposing that we are assessing the merits of two different systematizations of the *same* body of beliefs formulated in the *same* language. The significance of this restriction should be apparent from the fact that I have had to appeal on several occasions to the requirement that separate patterns cannot be fused in artificial ways, and that requirement assumes that the standards of artificiality are set by a *shared* language.

7.4 Comparative Unification and Scientific Change

However, it is a commonplace that scientific change may involve changes in belief and changes in language, both of which are justified on the grounds that the new beliefs and the new language have greater explanatory power than do the old. As a result, the principle of explanatory unification ought to be formulated so as to enable us to decide whether it would be reasonable to modify our scientific practice from $<L, K, E(K)>$ to $<L', K', E(K')>$ on grounds of attaining greater unification in our beliefs. I shall assume that such transitions have sometimes legitimately occurred in the history of science, for example, in the Darwinian revolution and in the birth of electromagnetic theory, and that something of the same kind is currently being envisaged by workers in particle physics. The difficulty is to allow for such changes without undermining our solutions to problems of asymmetry and irrelevance.

To see how such problems might re-emerge, reflect on the refined version of the asymmetry problem. I blocked the device of lumping many different dispositions in a single predicate by insisting that acceptable patterns must employ predicates that conform to the divisions made by the projectible predicates of our language. I have now admitted that linguistic change may be motivated by the desire to achieve explanatory gains. Hence, it would appear possible to defend a systematization whose basis included the *dispositional-shadow* pattern, by arguing that the gerrymandered predicate is a projectable predicate of the new language and recommending the linguistic change. As it stands there is no explanatory *gain* here—only the avoidance of explanatory loss—but once one is in the business of gerrymandering predicates, there will surely be ways of adding some further (unconnected) disjunct to the predicate that figures in the *dispositional-shadow* "pattern" and so generating at least one derivation with a conclusion that is not the conclusion of any derivation in $E(K)$.

Two separate issues arise here. First, we need to specify the conditions under which a systematization S of K provides a better unification of K than a systematization S' of K' does of K'. Second, we need to say when the fact that the best

systematization of K', $E(K')$, provides a better unification of K' than the best systematization of K, $E(K)$, does of K gives us reason to make the transition from $<L, K, E(K)>$ to $<L', K', E(K')>$. I shall not try to provide anything like a general account here, but will simply endeavor to offer partial conditions that seem to me to underlie important transitions in the history of science. To address the worry that allowing appeal to explanatory unification as a basis for scientific change, including linguistic change, reintroduces the asymmetry and irrelevance problems, I shall offer an obvious and intuitive suggestion. There is no force to the idea that switching to new language and/or new beliefs would enable one to have a more unified set of beliefs if there are principles that prohibit the kind of linguistic change or belief change envisaged. Thus, I propose that the search for unification of belief is conditional on principles that govern the modification of language and that rule on the acceptability of the proposed beliefs. It is easy to understand that there have to be some such principles, for otherwise unification could run riot over the deliverances of experience. My claim is that the principles are sufficiently powerful to preclude the maneuver envisaged in the previous paragraph in my resurrection of the asymmetry problem.

In considering the first issue, I shall restrict my attention to the special case in which the shift involves no explanatory loss. Assume that our principles (U), (C), (T), and (R) determine for two belief corpora K and K' unique best systematizations $E(K)$ and $E(K')$ respectively. Then the condition that the transition from $<L, K, E(K)>$ to $<L', K', E(K')>$ would involve no explanatory loss can be formulated as follows.

> For any statement that occurs as a conclusion of a derivation in $E(K)$ there is an extensionally isomorphic statement that occurs as a conclusion of a derivation in $E(K')$. Two statements are extensionally isomorphic just in case they have the same logical form and the nonlogical expressions at corresponding places refer to the same entities (objects in the case of names and sets in the case of predicates).

The notion of extensional isomorphism is introduced to permit the possibility that the shift from L to L' may involve refixings of reference, the kinds of changes that I have elsewhere (1978, 1982, 1983) taken to underlie the phenomenon of Kuhnian incommensurability. Thus Lavoisier would not accept Priestley's description of the phenomena that the phlogiston theory was used to explain, but he would be able to redescribe those phenomena in his own language. (Whether Lavoisier could also produce an explanatory derivation of a description of each of them, is, of course, a highly controversial matter.)[37]

When there is no explanatory loss (in the sense just characterized), it is not hard to formulate conditions that express the fact that there is a gain in explanatory unification in shifting from $<L, K, E(K)>$ to $<L', K', E(K')>$. Intuitively, there is explanatory gain if we would employ the same number of equally

stringent patterns to generate more consequences or fewer or more stringent patterns to generate the same consequences. So, presupposing that there has been no explanatory loss, we can formulate the idea that $E(K')$ unifies K' better than $E(K)$ unifies K as follows.

(C′) Suppose there is a one-one mapping f from the basis of $E(K')$ to the basis of $E(K)$ such that for each p in the basis of $E(K')$ p is at least as stringent as $f(p)$; and (i) f is an injection, or (ii) there is some p in the basis of $E(K')$ such that p is more stringent than $f(p)$, or (iii) there is some statement in the consequence set of $E(K')$ that is not extensionally isomorphic to any statement in the consequence set of $E(K)$. Then $E(K')$ unifies K' better than $E(K)$ unifies K.

To apply (C′) it is necessary to make sense of the notion of relative stringency for patterns that are not formulated in the same language. The way to make these comparisons can be suggested by some of the major examples in which large changes in science have been defended by appealing to the explanatory advantages of introducing the new language and the new theoretical claims. Consider Maxwellian electromagnetic theory. This supplies a variety of patterns for generating explanatory derivations within geometrical optics, the theory of diffraction, electrostatic interactions, and so forth. The old patterns of explanation are isomorphic to subpatterns of the new patterns, and unification is achieved because the same underlying pattern can be partially instantiated in different ways to generate patterns that were previously viewed as belonging to different fields. Similarly, in the Darwinian revolution, previously available patterns for explaining biogeographical distribution (such as they were) by appealing to the migrations of groups of organisms, were embedded within the Darwinian pattern with the initial premise describing an act of creation at some special center giving way to a derivation of a conclusion about the results of descent with modification.

I shall therefore propose that (C′) can be satisfied by meeting condition (i) if there is one (or more) pattern p of $E(K')$ such that there are at least two patterns of $E(K)$ that are extensionally isomorphic to subpatterns of p and if all other patterns of $E(K')$ – that is patterns that are not partially instantiable by isomorphs of patterns of $E(K)$ – are themselves extensionally isomorphic to patterns of $E(K)$. It seems to me likely that many episodes from the history of science will require comparisons of unifying power that are based on more subtle conceptions of relative stringency, but I shall not try to pursue this difficult issue further here.

Let us now turn to the second half of the problem that I posed above. Conceptual revision plainly occurs in the history of the sciences, leading to revocation of prior judgments about which predicates are projectable, which collections are natural, and which distinctions are artificial. If we allow the principle of explanatory unification full rein and suppose that advances in unifying power can be achieved through the strategy of embedding just described, then it would seem possible to make cheap explanatory improvements. Take any two unconnected

patterns in the basis of the current explanatory store. Form a new "pattern" in which both can be embedded. Typically, this new "pattern" will employ predicates that cut across the boundaries marked out by the projectable predicates of current language. Hence, *provided that the language is not itself amended*, the new "pattern" will be debarred by the ban on gerrymandered patterns. However that ban can be circumvented by proposing the adoption of a language in which the new predicates are taken to be projectable. As a result, if we allow that appeal to explanatory unification can serve as the basis of a defense of conceptual change – and I believe that such allowance is needed to cope with examples like that of the development of electromagnetic theory and of the Darwinian revolution – then we shall have to show how similar appeals will not justify spurious fusion of unconnected patterns.

I suggest that appeals to explanatory unification can underwrite transitions from $<L, K, E(K)>$ to $<L', K', E(K')>$ only subject to the proviso that the shifts from K to K' and the shifts from L to L' are defensible. This does not require that K and K', L and L' be identical – that would defeat the point of the current enterprise – but, roughly, that there are no strong arguments from the perspective of $<L, K, E(K)>$ against the shifts envisaged. Consider the simpler case first. Modification of K to K' may involve the addition of statements for which there was previously no positive evidence but which were not precluded by strong arguments from well-established principles of K (or conversely, such modification may involve abandoning statements in such a way that the prior view that there was evidence in favor of such statements is explained as illusory). When such modifications occur, or are proposed, I shall say that K is relatively *neutral* toward the changes. Contrast this with cases in which there are arguments using premises that are common both to K and to K' either against statements that would be added or in favor of statements that would be dropped. In the latter cases, where K is *negative* toward the changes, the proviso for the appeal to the increased unification is not met. However, when K is simply neutral toward the changes, the fact that the new corpus would permit greater unification of belief justifies the transition.

Darwin's argument in the *Origin* shows how the appeal to explanatory unification must be supplemented with a demonstration that the proviso is satisfied. As I have argued (Kitcher 1985a), the argument-strategy of the *Origin* consists in showing that certain kinds of modification of belief would enable a significant increase in unification. But Darwin is sensitive to the need for showing that accepted forms of reasoning *against* the modifications he would introduce – appeals to the uselessness of incipient complex structures, to the gappiness of the fossil record, and to the apparent stability of organisms – can be rebutted. If any of these major lines of objection had been left unanswered, then opponents could have responded that the theory of evolution by natural selection was an attractive fan-

tasy, promising the unification of belief at the cost of riding rough-shod over well-established facts.

The reception of the theory of continental drift in the 1920s and 1930s shows what happens when the proviso is not satisfied. Wegener and his early supporters could parallel *part* of Darwin's argument. They could show that if the assumptions about wandering continents were correct then considerable unification of geological, biogeographical, and paleometeorological beliefs could be obtained. Unfortunately, there was an apparently compelling argument against the possibility of moving continents. Wegener's own attempt to respond to the argument was unsuccessful, and a later effort at rebutting it (due to Arthur Holmes) was either unheeded or rejected by the geological community. In the absence of a demonstration that the proviso was satisfied, proponents of continental drift (such as du Toit) expanded the inventory of the advantages of unification in vain. Because geologists "knew" that it was impossible to move the continents, the unification described by Wegener and du Toit looked to them like an attractive fantasy that came to grief on well-established "facts" about the earth's crust.[38]

There is an analogous proviso about the modification of language. If we alter our language so as to change judgments about projectability, then we must respond to any existing arguments against the projectability of predicates that the new language takes to be projectable (or in favor of the projectability of predicates that the new language regards as unprojectable). Now in some cases, L is effectively neutral toward the projectability of new predicates: before the transition has been proposed, scientists have regarded certain phenomena as separate simply because they have seen nothing in common between them. But, in the transition, some predicate already viewed as projectable from the standpoint of L is taken to cover these phenomena. So, for example, in the birth of electromagnetic theory, both light and electromagnetic radiation are subsumed under the (projectable) predicate "transverse wave propagated with velocity c." Before the proposed subsumption, phenomena in the propagation of light would have been seen as unconnected with (what we see as analogous) phenomena involving electromagnetic effects, simply because light seemed to have nothing in common with either electricity or magnetism. Maxwell offered a common standpoint from which all these kinds of phenomena could be viewed—introducing a recognizably projectible predicate that would cover all of them—and there were no further negative arguments to be met.

However, in the artificial examples that threaten to resurrect the problems of asymmetry and irrelevance, there will be arguments against the projectability of the predicates that are employed in the new "patterns." Consider the trick of gerrymandering a new "pattern" by disjoining predicates. Let A, B, be disconnected predicates of L and suppose that L' proposes to treat "$Ax \vee Bx$" as a projectable predicate. Suppose that C, D are predicates such that "$(x)(Ax \supset Cx)$" and "$(x)(Bx \supset Dx)$" are generalizations accepted both in K and K', and such

that C, D, like A and B, are disconnected predicates of L. Now I take it that part of the reason for thinking that "$Ax \vee Bx$" is not a projectable predicate is that one could not confirm the generalization "$(x)((Ax \vee Bx) \supset (Cx \vee Dx))$" by observing a sample consisting of instances of A that are also instances of C. To make the proposed transition from L to L', it is not sufficient simply to declare that "$Ax \vee Bx$" is now to be counted as a projectable predicate. One will also have to answer arguments based on past inferential practice. Any such answer will, I believe, involve the modification of views about confirmation in such a way as to yield widespread changes in the corpus of beliefs—K' will have to differ from K in systematic ways—and some of the proposed changes will fall afoul of the proviso governing the modification of K to K'.

My tentative solution to the problem of how to allow for conceptual change based on an appeal to explanatory unification, while avoiding a revival of the asymmetry problem, is thus to insist that the appeal to advances in unification is contingent on the satisfaction of certain conditions. To *declare* that a predicate is projectable is not sufficient; one must respond to previously accepted reasons for not projecting the predicate. In doing so, it appears that there will have to be a large-scale revision of views about what confirms what. The change in inferential practice will lead to alterations in the corpus of beliefs that cannot be sustained in light of the arguments from premises that continue to be accepted.

We began this section with the worry that the idea of explanation as unification could not handle the problems of asymmetry and irrelevance. I have tried to show that, within the relatively simple context in which we consider a single language and a single corpus of beliefs, the unification criterion can be articulated to resolve major cases of these problems. With respect to the more complex context in which the possibility of change in belief and of conceptual change is allowed, matters are more complicated. I hope at least to have forestalled the complaint that the unification approach has no prospects of solving the apparently more recalcitrant asymmetry and irrelevance problems that may be posed within this context. Perhaps at this stage it is appropriate to see the burden of proof as resting with those who claim that the asymmetry and irrelevance problems doom any approach to explanation that does not explicitly invoke causal concepts in the analysis of explanation.

The discussion of appeals to unification and their role in the growth of scientific knowledge points to a more important project than resolving artificial problems of asymmetry and irrelevance. As I noted at the end of the last section, the demise of traditional foundationalist ideas about human knowledge leads to a reformulation of the question how our causal claims are justified. Ideally, we should show how our picture of the causal structure of the world has been built up, by tracing a chain of transitions leading from a hypothetical primitive state to the present. According to the approach to explanation as unification, principles like those presented in this section play a part in the dynamics of scientific knowl-

edge. Thorough vindication of the approach would require that those principles be articulated in greater detail, that they be embedded in a full set of principles for rational change in science, and that it be shown how such principles have been systematically operative in the growth of our scientific knowledge and, in particular, of our view of the causal structure of the world. Evidently, I am in no position to complete so massive a task, but the treatment sketched here offers some hope that the task might in principle be completed and that we can avoid the deep problems about causal knowledge that I tried to expose in the preceding section.

8. Metaphysical Issues

There is one final respect in which the two approaches to explanation need to be compared. Proponents of the causal approach to explanation can readily distinguish between acceptable and correct explanations. A true or correct explanation is one that identifies the causal structure that underlies the *explanandum* phenomenon. An acceptable explanation, for a person whose beliefs constitute a set K, identifies what it would be rational, given K, to take as the causal structure underlying the phenomenon. However, on the unification approach, the focus of the discussion has been on acceptable explanations. I have said what conditions must be met for a derivation to be an acceptable explanation of its conclusion relative to a belief corpus K, and (in the previous section) I have briefly discussed conditions governing the modification of belief and the modification of views about acceptable explanations. But I have not said what constitutes a true or correct explanation.

8.1 Correct Explanation

One obvious way of filling this gap is to say that d is a correct explanation of its conclusion just in case d belongs to $E(K)$ where K is the set of all true statements. But this will not do for several reasons. First, we would want K to be something like the set of all true statements in some particular language. Which language? Might it make a difference—particularly if the languages differ on their views about projectability of predicates? Second, if K includes *all* the true statements in a particular language, then it will include all true causal statements, all true statements about what explains what. So we could apparently shortcut our work by saying that d correctly explains its conclusion just in case a statement to that effect belongs to the set of all true English statements.

8.2 "What If the World Isn't Unified?"

I shall approach the problem of characterizing correct explanation more obliquely, starting with a difficulty that David Lewis has forcefully presented against views that take unification to be crucial to explanation. The difficulty can

be formulated in a short question: suppose the world isn't unified, what then? The idea is that proponents of the unification approach are committed to viewing factors as explanatorily relevant if they figure in a unified treatment of the phenomena. However, if the world is messy, then the factors that are causally relevant to the phenomena may be a motley collection and the account of explanatory structure may not reveal the causal structure. The unification approach thus runs the risk of providing an account of explanation on which giving explanations is divorced from tracing causes or of imputing to nature a priori a structure that nature may not have.

Let us elaborate the problem a little. The unification approach is apparently committed to claiming that some factor F is explanatorily relevant to a phenomenon P if there is a derivation belonging to the best unifying systematization of the complete set of truths about nature, a derivation in which a description of P is derived from premises that refer to F. To respond to the worries about the complete set of truths about nature that were raised at the beginning of the section, let us say that an *ideal Hume corpus* is a set of beliefs in a language whose primitive predicates are genuinely projectable (pick out genuine natural kinds) that includes all the true statements in that language that do not involve causal, explanatory, or counterfactual concepts. Then we can formulate one crucial premise of the objection as follows:

(1) F is explanatorily relevant to P, according to the criteria of explanatory unification, provided that there is a derivation of P that belongs to the best unifying systematization of an ideal Hume corpus such that there is a premise of the derivation in which reference to F is made.

A second important idea behind the objection is that explanations of at least some phenomena trace causes. Some proponents of the causal approach (Lewis, though not Salmon or Railton) hold that all singular explanation involves the tracing of causes. I have offered reasons for dissenting from this view (see section 3.2), but I have also suggested that it would be foolish for the unification approach to break the link between explanation and causation. Indeed, I have been emphasizing the idea (favored by Mill, Hempel, and many other empiricists) that causal notions are derivative from explanatory notions. Thus I am committed to

(2) If F is causally relevant to P then F is explanatorily relevant to P.

Now (1) and (2) are supposed to conflict with the genuine possibility that our world is messy and that there is a heterogeneous collection of basic causal factors. The idea of the possibility of a nonunified world can be presented as

(3) It is possible (and may, for all we know, be true) that there is a factor F that is causally relevant to some phenomenon P such that derivations of a description of P belonging to the best unifying systematization of an ideal Hume corpus would not contain any premise making reference to F.

(3) conjures up the vision of the obsessive unifier producing derivations of descriptions of *P* within a unified systematization of an ideal Hume corpus when the real causal factors underlying *P* are quite distinct from those that figure in the derivations. Unity is imposed where none is really to be found.

Now provided that the unification approach is conceived as offering (1) and (2) as necessary truths about explanatory and causal relevance, there is a genuine conflict among (1), (2), and (3). For (1) and (2) close off a possibility that (3) asserts to be open. One apparent way out would be to claim only that (1) and (2) formulate contingent conditions and that the unification approach is committed only to the idea that the world is actually unified. But this is not attractive, since the motivation for the unification approach rests on doubts about what explanatory relevance could be if it did not involve unification and about what causal relevance could be if it was not the projection of explanatory relevance. So it looks as though the approach must defend the *prima facie* implausible thesis that the world is necessarily unified—or, more exactly, that (3) is false.

Now the line of argument that I have been rehearsing not only raises difficulties for the attempt to combine the views that explanation consists in the unification of beliefs and that causal structure is explanatory structure; it also tells against efforts to co-opt some of the virtues of the unification approach for the causal approach. Salmon is sensitive to the fact that theoretical explanation often provides unification of the phenomena. In the final section of his (1984), he writes

> The ontic conception looks upon the world, to a large extent at least, as a black box whose workings we want to understand. Explanation involves laying bare the underlying mechanisms that connect the observable inputs to the observable outputs. We explain events by showing how they fit into the causal nexus. Since there seem to be a small number of fundamental causal mechanisms, and some extremely comprehensive laws that govern them, the ontic conception has as much right as the epistemic conception to take the unification of natural phenomena as a basic aspect of our comprehension of the world. *The unity lies in the pervasiveness of the underlying mechanisms* upon which we depend for explanation. (276)

It is clear from Salmon's earlier discussions that theoretical explanations are often defended on the grounds that they offer a unified treatment of phenomena previously regarded as diverse. But, once it is accepted that there is a causal structure of the world independent of our efforts to achieve a unified vision of it, then it is purely a contingent truth (supposing that it is indeed a truth) that there is a limited number of basic mechanisms. We are then forced to ask what evidence we may have for believing this supposed truth, and how we are able to appeal to it in our theoretical postulations.

On the causal approach, any use of a principle of explanatory unification to underwrite the revision of theory raises questions about how such a methodologi-

cal principle is to be justified. Salmon's contention that the ontic conception of explanation has as much right as the epistemic conception to take unification as a *basic* aspect of our comprehension of the world seems unfounded: for, on the version of the epistemic conception developed here (the unification approach) unification is *constitutive* of explanation, while on Salmon's version of the causal approach, unification is at best a *contingent concomitant* of the tracing of causal structure. The moral of Lewis's objection, elaborated in the tension among (1)–(3), seems to be that we must choose between two visions of what is central to explanation: the unification of phenomena or the identification of causal structure.

However, the discussion of Salmon's attempt to work ideas about unification into the causal approach indicates a different way of developing the metaphysics of explanation and of saving (1) and (2). What motivates our acceptance of (3) is the idea that there is an independent causal order, so that it is a purely contingent matter whether there are a few pervasive basic mechanisms or a motley assortment of fundamental causal factors. The heart of the unification approach is that we cannot make sense of the notion of a basic mechanism apart from the idea of a systematization of the world in which as many consequences as possible are traced to the action of as small a number of basic mechanisms as possible. In short, on the unification approach, the basic mechanisms must be those picked out in the best unifying systematization of our best beliefs, for if they were not so picked out then they would not be basic.

How can we defend the idea that (3) is false? Surely not by taking up Copernican ideas of the harmony of nature or Newtonian principles to the effect that nature "is wont to be simple and consonant to herself." Theological solutions that ultimately trace the necessary unity of nature to divine providence will not do. Instead, I recommend rejecting the idea that there are causal truths that are independent of our search for order in the phenomena. Taking a clue from Kant and Peirce, we adopt a different view of truth and correctness, and so solve the problem with which we began.[39]

8.3 Correct Explanation Again

Conceive of science as a sequence of practices, each of which is distinguished by a language, a body of belief, and a store of explanatory derivations. Imagine this sequence extending indefinitely forward into the future, and suppose that its development is guided by principles of rational transition, including the principles about unification outlined in the previous section. Instead of thinking that a language is ideal provided that its primitive predicates are projectable (or pick out genuine kinds) let us say that a predicate is correctly projectable if it is counted as a projectable predicate of the language of science in the limit as our sequence of practices develops according to the principles of rational transition. Similarly,

true statements are those that belong to the belief corpus of scientific practice in the limit of its development under the principles of rational transition. Finally, and most important for present purposes, *correct* explanations are those derivations that appear in the explanatory store in the limit of the rational development of scientific practice.

Is there a worrying circularity in the account that I have offered? Consider: the concept of projectability figured heavily in my discussion of the principles that govern rational modification of practices and the acceptability of argument patterns; yet now I am apparently planning to use those principles of rational modification to characterize the notion of a projectable predicate; surely something has gone wrong here. I reply that the circularity is only apparent. Recall that the enterprise of the sections up to and including 7 was to specify conditions on *acceptable* explanation. Thus, in focusing on the rational modification of scientific practices, I was concerned to understand what kinds of patterns of argument might be acceptable as explanatory against the background of judgments about natural kinds, judgments expressed in commitment to the projectability of various predicates. Hence, I used the concept of a predicate's *being accepted as projectable* to characterize that of an argument pattern's *being accepted as explanatory*. The dynamics of acceptance of predicates and of explanatory patterns was admittedly incomplete, but I believe that the principles governing rational modification of language, of beliefs, and of the explanatory store can be formulated independently of notions of correctness (of a predicate's being *actually projectable*, a belief's being *true*, and so forth). The proposal of the present section is to extend the envisaged account of rational modification of practices to an account of the corresponding concepts of correctness, by viewing correctness as that which is achieved in the ideal long run when the principles of rational modification are followed. There is no circularity here, although, as I would concede, there is plenty of work to be done before the program has been brought to completion.

Now we can respond to Lewis's worry about the unification approach and simultaneously articulate the idea that this version of the epistemic conception makes unification basic to explanation in a way that it is not basic to Salmon's version of the ontic conception. Given our new metaphysical perspective, (1) must be modified as follows

(1′) F is explanatorily relevant to P just in case, in the limit of the development of scientific practice under principles of rational transition, the explanatory store comes to contain a derivation in which a description of P is derived from premises, at least one of which makes reference to F. Among the principles of rational transition is a *static* principle of unification that directs that the explanatory store with respect to a body of belief K consists of those derivations that best unify K, and a *dynamic* principle that directs us to modify practice

so as to achieve advances in unification. The dynamic principle is subject to provisos, as indicated in the previous section.

The connection between explanatory relevance and causal relevance can go over unchanged. So we have

(2) If F is causally relevant to P then F is explanatorily relevant to P.

Finally, the claim that previously proved troublesome becomes

(3′) It is possible (and may, for all we know, be true) that there is a factor F that is causally relevant to some phenomenon P such that no derivation occurring in the explanatory store in the limit of scientific practice derives a description of P from premises that make reference to F.

The solution to Lewis's objection is to reject (3′) on the grounds that there is no sense to the notion of causal relevance independent of that of explanatory relevance and that there is no sense to the notion of explanatory relevance except that of figuring in the systematization of belief in the limit of scientific inquiry, as guided by the search for unification.

The difference between the role of unification on Salmon's approach and on mine can now be appreciated. It consists in the fact that Salmon would accept (3′) while I reject it. However, it is important to note that, although the problem posed by Lewis has been resolved, there is a sense in which the unification approach, in its new guise, can tolerate the possibility that the world might prove to be a messy place. Explanatory relevance emerges in the limit of our attempts to achieve a unified view of the world, but there is no a priori guarantee of how successful we shall be in achieving unification. Thus, while we try to make the phenomena as unified as we can, it is possible that there should be two different worlds, in one of which there were far fewer basic mechanisms than in the other. This would be reflected in the fact that the modification of scientific practice, guided in both worlds by the principles of rational transition, including both the static and dynamic principles of unification, produced in the one case a limit explanatory store whose basis contained only a few patterns and in the other an explanatory store with a far more prodigal basis. Hence, while my version of the unification approach makes it constitutive of explanatory relevance that there be no basic explanatory (or causal) mechanisms that are not captured in the limit of attempts to systematize our beliefs, it does *not* make it a necessary truth that this limit achieves any particular degree of unification. The vision of the obsessive unifier is dissolved.

8.4 Conclusions

As Railton clearly recognizes (see this volume) differences in views about scientific explanation connect to differences in metaphysics. The causal approach

is wedded to a strong version of realism in which the world is seen as having a structure independent of our efforts to systematize it. It should be no surprise that the metaphysical extras bring epistemological problems in their train (see section 6). I have been trying to show that we can make sense of scientific explanation and our view of the causal structure of nature without indulging in the metaphysics. The aim has been to develop a simple, and, I think, very powerful idea. The growth of science is driven in part by the desire for explanation, and to explain is to fit the phenomena into a unified picture insofar as we can. What emerges in the limit of this process is nothing less than the causal structure of the world.

Notes

1. This approach to pragmatic issues has been articulated with considerable sophistication by Peter Railton. See his (1981) and his unpublished doctoral dissertation.

2. Achinstein's theory of explanation, as presented in his (1983) is extremely complex. I believe that it ultimately suffers from the same general difficulty that I present below for van Fraassen. However, it is eminently possible that I have overlooked some subtle refinement that makes for a disanalogy between the two versions.

3. It would not be difficult to extend the account of ideal Hempelian questions to include questions that are answered by the provision of I-S explanantia. In this way we could mimic the entire Hempelian approach to explanation in the framework of the theory of why-questions.

4. As should now be evident, the second of the four problem-types that beset the Hempelian account assumes a derivative status. We may be able to manage the theory of explanation without a characterization of laws if we can distinguish the genuine relevance relations without invoking the notion of lawlike dependence. I shall articulate an approach below on which this possible strategy is attempted.

5. I am extremely grateful to Isaac Levi for raising the issue of the goals of a theory of explanation, by inquiring whether we can expect there to be a single set of relevance relations that applies for all sciences at all times.

6. The distinction between originating selection and maintaining selection is crucial, for the pressures need by no means be the same. A classic example is the development of wing feathers in birds, hypothesized to have been originally selected for their advantages in thermoregulation and subsequently maintained for their advantages in flight. The distinction is clearly drawn in Niko Tinbergen's celebrated account of the four "whys" of behavioral biology (see his 1968), and has been succinctly elaborated by Patrick Bateson (1982). Philosophical theories of teleology clearly need to use the contrast between maintaining selection and originating selection—see, for example, the treatments offered by Larry Wright (1976) and John Bigelow and Robert Pargetter (1987).

7. It should not, of course, be assumed that all functional/teleological questions are legitimized by the account that I have given. My concern is with the status of a class of why-questions, and it is compatible with my claims to suppose that the concepts of function and adaptation have been far too liberally applied in recent evolutionary theory. See, for example, Gould and Lewontin 1979, Gould and Vrba 1982, Kitcher 1985b, chapter 7, and Kitcher 1987a.

8. The distinction between global and local methodology is drawn in more detail in chapter 7 of (Kitcher 1983). It is only right to note that some scholars have challenged the idea that there is any very substantive global methodology. See, for example, Laudan 1984.

9. The version of the causal view that I present here is my own reconstruction of ideas put forward in seminar discussions by Railton and Salmon. Although I have some uneasiness about the mammoth causal histories that appear to be envisaged and about their status as "ideal" answers to "ideal" questions, I shall not press this line of objection. My main concerns about the causal approach lie elsewhere.

10. Plainly, this diverges from the Hempelian idea that the essence of an adequate explanation is that it show that the *explanandum* was to be expected. For discussion of this idea and motivation for an opposing viewpoint, see Salmon 1984, Railton 1978, and Railton's dissertation.

11. This apparent casualness can easily be understood by reviewing some of the literature on statistical explanation. In Salmon (1970), the Hempelian requirement of high-probability was replaced with a requirement of statistical relevance. Unfortunately, as it became clear during the 1970s, the notion of statistical relevance cannot do duty for the concept of causation – a *locus classicus* for the point is Cartwright (1979). Thus Salmon and others (e.g., Railton [1978, 1981], Humphreys [1981, 1982]) have explicitly introduced the notion of causation into the account of statistical explanation. In fact, the introduction can lead to more detailed models of statistical explanation than I consider here. See for example, the cited articles by Humphreys and Railton, and Salmon (1984), chapters 2 and 3.

12. Examples of this type have been suggested by Sylvain Bromberger in conversation and in his (1986).

13. Similar examples have been suggested by Peter Railton, who imagines explaining a person's failure to put a left-handed glove on the right hand.

14. This example has been used by Elliott Sober in discussions of causal explanation (1983). Although my treatment of it differs from Sober's in some matters of detail, we concur in viewing it as problematic for the causal approach.

15. We might think of the systematization approach as covering an entire family of proposals among which is that based on the view of systematization as unification. Since it appears to me that the latter view provides the best chances of success, I shall concentrate on it and ignore alternative possible lines of development.

16. See my (1986) for a reconstruction of Kant's views that tries to defend this attribution.

17. " . . . our total picture of nature is simplified via a reduction in the number of independent phenomena that we have to accept as ultimate" (Friedman 1974), 18. There is an interesting recapitulation here of T. H. Huxley's summary of Darwin's achievement. "In ultimate analysis everything is incomprehensible, and the whole object of science is simply to reduce the fundamental incomprehensibilities to the smallest possible number." (Huxley 1896) 165

18. I think it entirely possible that a different system of representation might articulate the idea of explanatory unification by employing the "same way of thinking again and again" in quite a different – and possibly more revealing – way than the notions from logic that I draw on here. Kenneth Schaffner has suggested to me that there is work in AI that can be deployed to provide the type of account I wish to give, and Paul Churchland has urged on me the advantages of connectionist approaches. I want to acknowledge explicitly that the adaptation of ideas about logical derivation may prove to be a ham-fisted way of developing the idea of explanatory unification. But, with a relatively developed account of a number of facets of explanation available, others may see how to streamline the machinery.

19. Carlson (1966) is an extremely valuable source for those who wish to interpolate. Extrapolation into the present is tricky, although Watson (1987) provides a helpful guide.

20. See the final section of Kitcher (1984), and Oster and Alberch (1982) for an account of *some* developmental complexities. Recent work on *Drosophila* only serves to underscore the point that epigenesis is hierarchically organized, and that changes near the top of the hierarchy can issue in a cascade of effects.

21. Since, as we shall see below, the primary Darwinian pattern of the more ambitious type is a pattern that traces the presence of prevalent traits to their selective advantages, the major debates concern the omnipresence and power of selection. Some critics (for example, Ho and Saunders 1984) contend that explanations in terms of natural selection are conceptually confused, even unintelligible. This radical claim, which lies behind much of the rhetoric about the demise of neo-Darwinism, does not seem to me to survive serious scrutiny, and is, I think, best viewed as the hypertrophied form of a sensible line of criticism that emphasizes the difficulties of identifying the workings of selection in particular cases. Those who pursue this more moderate line are quite willing to allow that explanations instantiating the primary Darwinian pattern can be given in particular cases, but they believe that many biologists are carried away by enthusiasm for paradigm cases in which the primary Darwinian pattern has proved successful, so that they overlook alternative forms of explanation in studies

where rival possibilities deserve to be taken seriously. Cautionary notes of this kind can be found in (Gould and Lewontin 1979) and in (Bateson 1982). I have tried to articulate the moderate criticism in my (1985b).

22. Aficionados of sorites problems will recognize that the use of induction in the context of reasoning about "almost all" members of a group is dangerous. The deficiency could be made up by reformulating the reasoning in probabilistic terms, or simply by adding a premise to the effect that lapses from universality are noncumulative. In the interests of preserving something that is fairly close to Darwinian arguments, I shall avoid such niceties here.

23. This picture of the structure of contemporary evolutionary theory is akin to that provided by Sober (1984). Ecology can be seen as providing principles about the sources of the factors that are cited in the equations of population genetics. Chapters 2 and 3 of Kitcher (1985b) develop the picture in the context of evolutionary/functional studies of animal behavior.

24. The version of the deductivist gambit presented here is quite close to that offered in Papineau (1985). Besides Papineau, Levi has contributed important ideas to the defense of deductive chauvinism. See Levi forthcoming.

25. Ironically, this is very close to Hempel's paradigm example of singular explanation in his 1965 (232). The similarity reinforces the idea that explanation is implicitly global, the point from which Friedman (1974) began his seminal discussion of unification, and which I am trying to develop further here.

26. Here I use standard talk of determinism and indeterminism to distinguish QM from other parts of science. John Earman (1986) has shown in lucid and thorough detail how loose some of this talk is, but the inexactness will not affect the present arguments.

27. Behind the dilemma enunciated here there are significant epistemological presuppositions. I shall consider a possible way of altering the epistemological framework at the end of this section (see 6.5).

28. This point will be elaborated briefly in 7.3.3.

29. Intuitively, we have no interest in an initial interaction that produced a mark in a process that transmitted that mark and interacted at the *explanandum* event to produce some completely irrelevant property. We are concerned with the interaction that gave rise to the very property whose presence is to be explained, and whatever other markings and transmittings are going on at the point of the *explanandum* event or along the line that constitutes its causal history can be ignored.

30. See Lewis (1973), both for an elegant statement of a counterfactual theory of causation and for a survey of difficult cases. Loeb (1974) endeavors to cope with the problem of overdetermination.

31. Others have seen that a relative of Benacerraf's dilemma applies to modal discourse. See, for example, Mondadori and Morton (1976). Unfortunately, the significance of the point does not appear to be widely appreciated.

32. Some accounts of causation, for example those of Cartwright (1979) and Sober (1984), provide recursion clauses for enabling some causal judgments to be generated from others. Valuable as this is, it does not explain the basis of our practice of causal justification.

33. This line of attack was forcefully presented in seminar discussions by Paul Humphreys, who maintains that it can be developed to overcome the obstacles to Salmon's program (and programs, like Humphreys's own, which are akin to Salmon's) that I have presented here.

34. This can easily occur in cases of explanation extension. See the examples from section 4.6.

35. Notice that derivations that systematically contain idle clauses are not so clearly nonexplanatory as the kind of irrelevant derivation with which we began. It seems to me that this is because the unwanted mini-derivations are viewed as giving us information about the structure of a full, ideal, derivation, and the natural implication is that the properties picked out in the premises will play a key role. Once we see that these properties are inessential, and that predicates referring to them figure throughout all our derivations, then we may feel that cluttering up the explanatory store does nothing more than add a harmless irrelevancy. The resultant derivations are untidy, but I think that there is reason to argue about whether they should be counted as nonexplanatory.

36. For this example, it is important to recognize that the *origin-and-development* pattern must allow for explanatory derivations in which we appeal to general constraints that keep a system close to an equilibrium state throughout its career. See the discussion of the sex-ratio example in 3.2.

37. Numerous scholars have contended that examples like this involve what has come to be known as "Kuhn-loss." In the present case, it is sometimes suggested that Priestley's phlogiston chemistry could explain something that Lavoisier could not, namely what all the metals have in common. I shall take no stand on this vexed question here.

38. For a brief, comprehensible account of the main ideas and the course of the debate, see Hallam (1973). Comparing the arguments of du Toit with the responses of the leading opponents of continental drift in the 1930s and 1940s makes it quite apparent that the appeal to unifying power had to be subject to provisos. However, in the spirit of Kuhn (1977), it seems that scientists sometimes differ with respect to the advantages of unifying and the severity of the arguments against unifying. In consequence, there are sources of diversity within the scientific community, and this, I would contend, is a good thing from the point of view of the *community's* enterprise.

39. The view that I shall develop from hints of Kant and Peirce is more closely connected with Kantian texts in my (1986). It also has affinities with ideas presented in Rescher (1970) and in the recent writings of Hilary Putnam.

References

Achinstein, Peter. 1983. *The Nature of Explanation*. New York: Oxford University Press.

Bateson, P. P. G. 1982. Behavioural Development and Evolutionary Processes. In *Current Problems in Sociobiology*, eds. King's College Sociobiology Group. Cambridge: The University Press.

Benacerraf, Paul. 1973. Mathematical Truth, *Journal of Philosophy*, 70: 661–79

Bigelow, John, and Pargetter, Robert. 1987. Functions, *Journal of Philosophy*, 84: 181–96.

Bromberger, Sylvain. 1963. A Theory about the Theory of Theory and about the Theory of Theories. In *Philosophy of Science: The Delaware Seminar*, ed. W. L. Reese. New York: John Wiley.

——. 1966. Why-Questions. In *Mind and Cosmos*, ed. R. Colodny. Pittsburgh: University of Pittsburgh Press.

——. 1986. On Pragmatic and Scientific Explanation: Comments on Achinstein's and Salmon's Papers. In *PSA 1984*, Volume II, eds. Peter Asquith and Philip Kitcher. East Lansing: Philosophy of Science Association

Carlson, E. O. 1966. *The Gene: A Critical History*. Philadelphia: Saunders.

Cartwright, Nancy. 1979. Causal Laws and Effective Strategies, *Noûs*, 13: 419–37. (Reprinted in Cartwright 1983.)

——. 1983 *How the Laws of Physics Lie*. Oxford: the University Press

Coffa, J. Alberto. 1974. Hempel's Ambiguity, *Synthèse* 28: 141–64.

Dretske, Fred. 1973. Contrastive Statements, *Philosophical Review*, 82: 411–37.

Earman, John. 1986. *A Primer on Determinism*. Dordrecht, Holland: D. Reidel.

Feigl, Herbert. 1970. The 'Orthodox' View of Theories: Remarks in Defense as well as Critique. In *Minnesota Studies in the Philosophy of Science*, Volume IV, eds. M. Radner and S. Winokur. Minneapolis: University of Minnesota Press.

Fisher, R. A. 1931. *The Genetical Theory of Natural Selection*. Oxford: the University Press. (Second Edition, New York: Dover, 1958)

Friedman, Michael. 1974. Explanation and Scientific Understanding, *Journal of Philosophy*, 71: 5–19.

Garfinkel, Alan. 1981. *Forms of Explanation*. New Haven: Yale University Press.

Geach, Peter T. 1969. *God and the Soul*. London: Routledge and Kegan Paul.

Goodman, Nelson. 1956. *Fact, Fiction, and Forecast*. Indianapolis: Bobbs-Merrill.

Gould, S. J., and Lewontin, R. C. 1979. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. Reprinted in *Conceptual Issues in Evolutionary Biology*, ed. Elliott Sober. Cambridge MA.: Bradford Books/MIT Press, 1983.

Gould, S. J., and Vrba, E. 1982. Exaptation: A Missing Term in the Science of Form, *Paleobiology*, 8: 4–15.

Hallam, Anthony. 1973. *A Revolution in the Earth Sciences*. Oxford: the University Press.

Hempel, C. G. 1965. *Aspects of Scientific Explanation*. New York: Free Press.

Hempel, C. G., and Oppenheim, P. 1948. Studies in the Logic of Explanation. Chapter 9 of Hempel (1965).

Ho, Mae-Wan, and Saunders, Peter T. 1984. *Beyond Neo-Darwinism: An Introduction to the New Evolutionary Paradigm*. New York: Academic Press.
Horwich, Paul. 1987. *Asymmetries in Time*. Cambridge MA: MIT Press/Bradford Books.
Humphreys, Paul. 1981. Aleatory Explanations, *Synthèse*, 48: 225–32
——. 1982. Aleatory Explanations Expanded. In *PSA 1982*, eds. Peter Asquith and Thomas Nickles. East Lansing: Philosophy of Science Association.
Huxley, Thomas Henry. 1896. *Darwiniana*. New York: Appleton.
Jeffrey, Richard. 1969. Statistical Explanation vs. Statistical Inference. In *Essays in Honor of Carl G. Hempel*, ed. Nicholas Rescher. Dordrecht, Holland: D. Reidel.
Kim, Jaegwon. 1974. Noncausal Connections, *Noûs*, 8: 41–52
Kitcher, Philip. 1975. Bolzano's Ideal of Algebraic Analysis, *Studies in the History and Philosophy of Science* 6: 229–71.
——. 1976. Explanation, Conjunction, and Unification, *Journal of Philosophy* 73: 207–12.
——. 1978. Theories, Theorists, and Theoretical Change, *Philosophical Review*, 87: 519–47.
——. 1981. Explanatory Unification, *Philosophy of Science*, 48: 507–31.
——. 1982. Genes, *British Journal for the Philosophy of Science*, 33: 337–59
——. 1983. *The Nature of Mathematical Knowledge*. New York: Oxford University Press.
——. 1984. 1953 And All That. A Tale of Two Sciences, *Philosophical Review*, 93: 335–73.
——. 1985a. Darwin's Achievement. In *Reason and Rationality in Science*, ed. N. Rescher. Washington: University Press of America.
——. 1985b. *Vaulting Ambition: Sociobiology and the Quest for Human Nature*. Cambridge MA.: MIT Press.
——. 1985c. Two Approaches to Explanation, *Journal of Philosophy*, 82: 632–39.
——. 1986. Projecting the Order of Nature. In *Kant's Philosophy of Physical Science*, ed. Robert Butts. Dordrecht: D. Reidel.
——. 1987a. Why Not The Best? In *The Latest on the Best: Essays on Optimality and Evolution*, ed. John Dupre. Cambridge MA.: Bradford Books/MIT Press.
——. 1987b. Mathematical Naturalism. In *Essays in the History and Philosophy of Modern Mathematics*, eds. William Aspray and Philip Kitcher. Minneapolis: University of Minnesota Press, 293–325.
——. 1988. Mathematical Progress. To appear in *Revue Internationale de Philosophie*.
——, and Salmon, Wesley. 1987. Van Fraassen on Explanation, *Journal of Philosophy*, 84: 315–30.
Kuhn, Thomas S. 1970. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.
——. 1977. Objectivity, Value Judgment, and Theory Choice. In *The Essential Tension*. Chicago: University of Chicago Press.
Kyburg, Henry. 1965. Comment, *Philosophy of Science*, 35: 147–51.
Laudan, Larry. 1984. *Science and Values*. Berkeley: University of California Press.
Levi, Isaac. (forthcoming) Four Types of Statistical Explanation. To appear in *Probabilistic Causality*, eds. W. Harper and B. Skyrms. Dordrecht, Holland: D. Reidel.
Lewis, David. 1973. Causation, *Journal of Philosophy*, 70: 556–67.
——. 1974. *Counterfactuals*. Oxford: Blackwell.
——. 1976. The Paradoxes of Time-Travel, *American Philosophical Quarterly*, 13: 145–52.
Loeb, Louis. 1974. Causal Theories and Causal Overdetermination, *Journal of Philosophy*, 71: 525–44.
Mondadori, Fabrizio, and Morton, Adam. 1976. Modal Realism: The Poisoned Pawn, *Philosophical Review*, 85: 3–20.
Oster, George, and Alberch, Pere. 1982. Evolution and Bifurcation of Developmental Programs, *Evolution*, 36: 444–59.
Papineau, David. 1985. Probabilities and Causes, *Journal of Philosophy*, 82: 57–74.
Pauling, Linus. 1960. *The Nature of the Chemical Bond* (Third Edition), Ithaca: Cornell University Press.
Railton, Peter. 1978. A Deductive-Nomological Model of Probabilistic Explanation, *Philosophy of Science*, 45: 206–26.
——. 1981. Probability, Explanation, and Information, *Synthèse*, 48: 233–56.

Rescher, Nicholas. 1970. Lawfulness as Mind-Dependent. In *Essays in Honor of Carl G. Hempel*, ed. N. Rescher. Dordrecht, Holland: D. Reidel.

Salmon, Wesley. 1970. Statistical Explanation. In *The Nature and Function of Scientific Theories*, ed. R. Colodny. Pittsburgh: University of Pittsburgh Press.

——. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Scriven, Michael. 1959. Definitions, Explanations, and Theories. In *Minnesota Studies in the Philosophy of Science* Volume II, eds. H. Feigl, M. Scriven, and G. Maxwell. Minneapolis: University of Minnesota Press.

——. 1962. Explanations, Predictions, and Laws. In *Minnesota Studies in the Philosophy of Science* Volume III, eds. H. Feigl and G. Maxwell. Minneapolis: University of Minnesota Press.

——. 1963. The Temporal Asymmetry between Explanations and Predictions. In *Philosophy of Science. The Delaware Seminar*, Volume I, ed. B. Baumrin. New York: John Wiley.

Sober, Elliott. 1983. Equilibrium Explanation, *Philosophical Studies*, 43: 201–10.

——. 1984. *The Nature of Selection*. Cambridge, MA: Bradford Books/MIT Press.

Stalnaker, Robert. 1968. A Theory of Conditionals. In *Studies in Logical Theory*, ed. N. Rescher. Oxford: Blackwell.

Steiner, Mark. 1978. Mathematical Explanation, *Philosophical Studies* 34: 135–51.

Tinbergen, Niko. 1968. On War and Peace in Animals and Man. Reprinted in *The Sociobiology Debate*, ed. Arthur Caplan. New York: Harper and Row, 1978.

van Fraassen, Bas. 1980. *The Scientific Image*. Oxford: the University Press.

Watson, J. D. 1987. *Molecular Biology of the Gene* (Fourth Edition). San Francisco: Benjamin.

Wright, Larry. 1976. *Teleological Explanation*. Berkeley: University of California Press.