CHAPTER 1

Cognitive Strategies and Scientific Discovery

> We have come, or are coming, at last to the end of this epoch, the epoch presided over by the concepts of Newtonian cosmology and Newtonian method. We are in the midst of a new philosophical revolution, a revolution in which, indeed, the new physics too has had due influence, but a revolution founded squarely on the disciplines concerned with life: on biology, psychology, sociology, history, even theology and art criticism.
>
> —M. Grene 1974

1. Rationalizing Scientific Discovery

Logical positivism drew a principled and sharp distinction between the context of justification and that of discovery. According to this view, the empirical evaluation of scientific theories could be submitted to logical analysis, with the goal of specifying the conditions under which a theory would be confirmed or disconfirmed. Theories were accordingly given an axiomatic rendering. The fundamental laws provided the axioms in terms of which all else was explained. The theorems were either nonfundamental laws or observational claims. Scientific discovery, by way of contrast, seemed to be far less congenial to logical analysis and was therefore shunted off as a "merely" psychological matter having no significant consequences for a theory of scientific rationality. Hans Reichenbach captured the spirit of the times in his classic work, *The Rise of Scientific Philosophy:*

> The act of discovery escapes logical analysis; there are no logical rules in terms of which a 'discovery machine' could be constructed that would take over the creative function of the genius. But it is not the logician's task to account for scientific discoveries; all he can do is to analyze the relation between given facts and a theory presented to him with the claim that it explains these facts. In other words, logic is concerned with the context of justification. (1951, p. 231)

Discovery correspondingly was seen as a mystical process; justification alone was susceptible to logical analysis. Even those like Karl Popper, who dogmatically rejected the idea that scientific theories could ever be verified, held on to the distinction between what could be logically analyzed and what had to be left to psychology. In the opening chapter of *The Logic of Scientific Discovery*, Popper writes:

> The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it. The question how it happens that a new idea occurs to a man . . . may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge. This latter is concerned not with *questions of fact*, . . . but only with questions of *justification or validity*. (1959, p. 31)

Scientific discovery is, ironically, left aside as a topic for psychology, but not for logic.[1]

More recent philosophy of science, together with the branches of cognitive science, has returned to the investigation of discovery, recognizing that the distinction between discovery and justification is artificial and that there are rational or 'logical' considerations guiding both processes (for example, see Darden 1980 and 1991; Giere 1979, 1988; Nickles 1980, 1982, 1987a, 1987b; Simon 1977; and Thagard 1988 and 1992). Pressure against any firm distinction between justification and discovery correspondingly comes from two directions. One challenges the purity of a 'logic' of justification; treating justification in abstraction from psychological processes, social context, and historical setting also abstracts from the scientific practice that gives it life. The other demystifies the process of discovery; once recognized for the problem-solving process it is, discovery too is animated by scientific practice. If discovery and justification are not quite one, they are at least reflections of a single activity that finds its expression in scientific practice—though not, of course, *only* in scientific practice.

Positivistic accounts not only ignore discovery, but also see justification in exclusively empirical terms. Theory confirmation and disconfirmation involve more than simple considerations of empirical adequacy—empirical data has unclear relevance in the absence of an appropriate theory.[2] Scientists were uncertain of the significance of, for example, the data on the relative frequencies of neucleotide bases prior to the recognition of complementary binding in Watson's and Crick's double helix. And Mendel's hybridization ratios did not have the significance they have for us before the study of genetics was divorced from embryology and development in the early decades of the twentieth century. Empirical data, moreover, may be regarded as irrelevant even when it cannot be reconciled with a favored theory. It is well known, for example, that the deviations in the orbit of Mercury were observed and recognized, but largely ignored, before the advent of the theory of relativity; similarly, despite considerable evidence for maternal inheritance, cytoplasmic inheritance found no significant place within a theory committed to nuclear inheritance, and evidence of maternal inheritance was dismissed as irrelevant rather than disconfirmatory.[3] Empirical data also may be reconciled with a theory by a variety of apparently ad hoc devices, as discontinuities in the fossil record were dismissed in nineteenth-century evolutionary debates. Empirical constraints are simply not enough to explain theoretical relevance. If we are to understand confirmation or disconfirmation as it occurs in scientific practice, we must seek additional constraints from other sources. Likewise, while empirical constraints are important to discovery, they are not sufficient here either. A model capable of accommodating either confirmation or discovery must incorporate empirical constraints, but it must incorporate further constraints as well.

There is a wide range of constraints that are potentially relevant, including some that are broadly social, historical, or technological. That psychological constraints are also at least *part* of what is needed for a theory of justification is suggested by the observation that logicist accounts of justification operate with unrealistic assumptions concerning what human beings can do. For example, despite considerable Bayesian literature on confirmation, it is reasonably clear that humans systematically undervalue prior probability, to the point of ignoring it altogether, even when given "data" that is entirely worthless. In psychological experiments, humans violate fundamental rules of probability, ignore base rates, and neglect questions of sample size in making predictions (Kahneman and Tversky 1972, 1973; Tversky and Kahneman 1974), fail to seek disconfirmatory evidence even when it is readily obtained, do not use available disconfirming evidence, and ignore alternative explanations of the available data (Johnson-Laird and Wason 1970; Wason 1966). Though one might wish it were not so, scientists do little better (see Faust 1984). Clinical prediction has long been known to be less reliable than actuarial judgment: in a variety of areas, from neuropsychology to psychiatry, it is clear that we can get more reliable prediction and diagnosis by using mechanical means than by relying on human judgment. Nonetheless, there is a pervasive tendency to overestimate human judgment's usefulness. For example, clinicians given results generated by actuarial means will place undue reliance on their own intuitive judgment, discounting the statistical data. Such discretionary judgment reduces reliability.[4] Moreover, as Amos Tversky and Daniel Kahneman (1971) argue, the situation is not appreciably improved by turning to individuals trained in statistics (for some mitigating considerations, see Nisbett 1988). Even statistically sophisticated psychologists adopt sample sizes in replication studies that are unreasonably low, and *then* underestimate the significance of the results. If formalist theories prescribe procedures that are fundamentally different from those constitutive of human reasoning, it should be little surprise when their prescriptions diverge from what we actually do. Yet science is a human enterprise; if the problem is understanding the processes of human confirmation and human discovery, we must evidently look to human reasoning.

We propose to treat scientific change, scientific discovery, and scientific rationality in a way that is simultaneously developmental and psychological. Our approach is *developmental* in emphasizing the dynamics of theoretical change rather than its statics and in embracing the historical contingency of the resulting theories. In positivist and neopositivist models of theories, questions of the dynamics of theories are treated as derivative from their static descriptions. Theories are taken as axiomatized systems. Convergence and commensurability are the dynamic variables, and they, in turn, reduce to similarity in theoretical commitments. The neopositivist emphasis is on theories abstracted from their development. By contrast, we seek the laws of motion. It is in modification and change, in theories in disequilibrium and flux, that we can hope to reveal the cognitive strategies that drive scientific change and are constitutive of scientific rationality. Classical models of theories were self-consciously ahistorical, divorcing contemporary content from its origins. Our emphasis on dynamics leaves no room for such luxury. To shift the image from the physical sciences to the biological, we look to the ontogeny of theories. In the ontogeny of an organism, an earlier developmental stage may exemplify structures that have no counterparts in the adult organism. Nevertheless, these stages are crucial to development, and the resulting structures cannot be understood in abstraction from their sources in earlier developmental stages: the obsolete earlier stages make the later stages possible. In parallel fashion, later theoretical developments in a scientific discipline may have little in common with their precursors. Nevertheless, these early stages may be crucial to the emerging discipline, and it is not generally possible to understand the theoretical commitments at a given time in abstraction from the commitments of precursors. In the ontogeny of theories, no less than in the ontogeny of organisms, historical contingency is the consequence of developmental priority (cf. Wimsatt 1986b).

Our approach is *psychological* in emphasizing that theoretical development is in part an expression of human cognitive style, a consequence of the typical strategies with which human beings attack problems, and the cognitive limitations that make these strategies necessary. Discovery and justification alike should be understood in psychological terms. A model of scientific discovery should treat scientific problem-solving as a special case of human problem-solving. As such, it is subject to all the contingencies of human problem-solving, both internal and external, psychological and social. The approach to theoretical, explanatory, and empirical problems posed by scientific investigation is determined, in part at least, by cognitive capacities and limitations. Humans do *not* perform exhaustive searches, even when the problems faced are simple enough to make it feasible. Humans do *not* operate with axiomatized structures, searching the consequence set for confirmation and disconfirmation. Explanatory models, simplified schemata, and the assimilation of complex phenomena to these prototypes is typical of human cognitive style (cf. Cartwright 1983; Johnson-Laird 1983; Laymon 1980, 1982, 1989; McCauley 1985; Richardson 1986b, c). The process is one of adjustment and pattern matching, rather than analysis and deduction. Search is selective and limited, guided by heuristics, local in its application, and biased in its outcome. To psychologize the treatment of scientific change is to take these issues of cognitive style seriously. The theorist is, after all, important to the theory.

Our enterprise is, thus, not only psychological, but self-consciously *psychologistic*. This does not mean it is wholly descriptive. We operate on the assumptions that psychological constraints are important in understanding scientific change, confirmation, and discovery, and that the sorts of explanatory strategies employed in scientific problem-solving are analogous to the strategies employed elsewhere in human problem-solving. We are concerned primarily with how an explanatory task is seen and how the problem is represented; with how the representation of the problem can influence the character of the resulting theories, and how both are affected by the conceptual and technical tools available in addressing the problems; with what explanatory strategies were adopted, what reasons there were for adopting these strategies, and how explanatory strategies are modified in the face of a new representation of the problem. These are not questions of individual eccentricities or personal quirks; they are applications of problem-solving strategies that have a wide scope, and they are intended as faithful approximations of human cognition. If these explanatory strategies were not common, our own descriptions might be charged with being no more than post hoc descriptions of cases. The fact that they are robustly represented in human cognition, however, constitutes the case for their reality.

The model we will present is itself an approximation, or, more accurately, an idealization—a fragment of a theory of scientific competence. It is a *fragment* because the explanatory strategy on which we will focus is only one among many and is not well suited for all cases. As we will explain in the next chapter, the strategy we will discuss involves *decomposition and localization*—that is, a hierarchical analysis into functional components with specific functions. This strategy is particularly well suited to problems that are relatively ill defined, problems in which neither the criteria for a correct solution nor the means for attaining it are clear. In cases in which the explanatory domain is better defined, alternative strategies may be better suited to the task. There are varying constraints on what constitutes an adequate solution to problems. Experimental methodology and technology are generally tied to these constraints, both by determining what can be answered and by reflecting current theories.

One of the strengths of decomposition and localization as a scientific strategy is that it facilitates an increasingly realistic representation of the explanatory domain, even when the initial representation is seriously distorted: failures of localization can be as revealing as successes. What we offer is a model of scientific *competence* because it is an idealized description, largely abstracting on the one hand from individual eccentricities, and on the other from the vicissitudes of social circumstance. No doubt, both factors will prove essential to any full and descriptively adequate account of scientific change. Galileo's ability to grind lenses surely affected the fate of his theory, as did the need for his expertise in ballistics. Similarly, the opportunities afforded to Pasteur by the French wine industry were no less important to the line of his research than was his role in the French Academy or his expertise with glassware.

We will, thus, focus on *one* constraint on scientific development and discovery. It is a cognitive dimension, but one among many relevant constraints, including a broad range that can be classed as empirical, practical, technological, historical, social, and theoretical. Scientific development, accordingly, becomes a process of simultaneously satisfying a host of constraints. By focusing our attention on one, we will see that others stand out, as if in relief. We will, moreover, discuss only one set of heuristics that might be deployed. As we have suggested, these heuristics are useful when confronted with ill-defined problems in a complex domain.

We propose to use historical cases in detailing the significance of psychological constraints on the development of theories. There are a number of methodological perils in the use of historical case studies, or in what Kevin Kelly and Clark Glymour (1990) derisively call a "retreat to history." One obvious problem is that historical cases can be selected simply because they support some favored theme or set of themes. The use of historical cases then degenerates into elaborate proof-texting and is of little use at all. This problem can be mitigated by using independently plausible and robustly characterized psychological heuristics. For example, Miriam Solomon (1992) has shown that cognitive heuristics proposed and investigated by Tversky and Kahneman can be used to explain a variety of historical features in the development of continental drift in the mid-1900s. Salience, or vividness, of evidence is enhanced if it is concrete, or a matter of personal experience. Such evidence is given a disproportionate weight in comparison to more abstract, and more reliable, statistical evidence. An anecdote from a friend about her car, and the many trips to the shop for its myriad problems, will often outweigh the cold and dry statistical evidence on reliability and repair records provided by *Consumer Reports*. Solomon shows that salience can be used to explain, for example, why scientists with a personal exposure to the biota and geology of the southern hemisphere were more likely to accept drift than were their contemporaries who worked primarily with northern hemisphere data. The same data was available to all. The main differences lie in the way the data was weighted, and this is what Solomon explains in terms of salience, understood as a cognitive heuristic. Solomon does not offer the case study as evidence for the reality of salience. The experimental work of Tversky, Kahneman, and their collaborators has amply demonstrated the role and importance of salience in human cognition. Solomon uses her research, rather, to explain patterns of acceptance in the geological community. By using a mechanism that is independently motivated, Solomon avoids any charge of proof-texting.

We too begin with a heuristic that is independently motivated, though it is lacking the kind of systematic and detailed experimental support provided for salience by Tversky and Kahneman, and we also examine its role in a variety of historical cases. This allows us to explain a number of patterns in the use of evidence and the development of explanatory models, and also provides us with a kind of null hypothesis—a backdrop of expectations against which we may see the operation of other constraints on the development of theories. A parallel might help: It is clear that values affect scientific practice in a variety of ways. The role of social values, or ideological commitments, is particularly evident when theoretical decisions or the assessment of evidence deviates from what would otherwise be normatively appropriate (cf. Richardson 1984). This has led some philosophers to suggest that ideological commitments are present or relevant *only when decisions are otherwise unreasonable*. Larry Laudan, for example, describes this as the *arationality assumption*, the view that sociological and ideological explanations are not necessary when there is an adequate rationalization (1977, pp. 201ff.). This view, however, is mistaken (cf. Longino 1990; Martin 1989): a deviation from what is rationally required *does* make the influence of values more obvious and more readily detectable, but their influence and importance are no less real in cases where they are less obvious. We can use a psychological null hypothesis to detect the presence of other influences on theory development, without holding that these influences are absent unless they are readily detected.[5] To assume that detecting *if* a heuristic is present is tantamount to determining *when* it is present is an egregious and unwarranted abuse of the method.

We thus propose to use historical cases, examining the extent to which we can explain developments and patterns in the history in terms of psychological influences. Through this method we will also see that some developments and patterns cannot be so explained, and we will use such instances as a means for discerning other influences on the development of theories. This will allow us to get some grasp of the range of influences and the importance of psychological constraints on scientific development. We do not retreat to history. We embrace it.

This does not mean we have abandoned the hope for a normative theory of scientific discovery or scientific rationality. Simultaneously psychologizing the account of scientific rationality and insisting on its historical, developmental character does leave us straddling an issue that has loomed large in philosophical thinking about science and philosophy of science. Philosophers of science, like most epistemologists, have presented their endeavors as normative. The task has been thought of as one of prescribing or proscribing what an adequate scientific explanation is, and what is necessary for scientific justification, independent of limitations on decision making or decision makers. The techniques of logical analysis were once supposed to be sufficient to settle these issues. We abandon the a prioristic aspirations and look instead to scientific practice and human reason. We insist on a realistic theory of *human* rationality, and correspondingly on a realistic theory of *scientific* rationality; yet, the enterprise is still, in part at least, normative (cf. Goldman 1986). We propose to reconstitute the normative enterprise rather than reject it.

It is uncontroversial to point out that we cannot simply decide what it is right to do by noting what is actually done. The gap between *is* and *ought* is not this narrow. That apparently leaves us with few options. One is simply to settle for a description of actual practice. We then abandon the search for normative guidelines, or treat normative standards as themselves constituted by particular theories or research traditions. To do this is to yield to a naive historical relativism. Another option is to adopt utopian norms, divorced from the historical episodes, and to ignore the theorist and the tradition in favor of the theory. Just as we may study how clinicians approach medical diagnosis, we may study how scientists approach the task of theory construction and validation; just as we may contrast the practice of clinicians with the formal and analytical actuarial methods that are "normatively appropriate," we may choose to contrast the practice of science with some "normatively appropriate" standard derived from logic or probability theory. Much scientific reasoning—including much of the best—will then violate the norms. To do this is to yield to historical idealism. We seek to forge an alternative between these extremes.

By recasting the normative endeavor to take account of the realities of scientific inquiry and scientific inquirers, we defend a *naturalized* and *humanized* theory of scientific rationality. By describing the strategies scientists use, and simultaneously seeing their limits, we leave room for a normative assessment. We can evaluate actual performance and different human strategies by identifying contexts where they succeed and where they fail, and by seeing which constraints are sufficient and which are not. Such a naturalized theory avoids the Scylla of historical idealism by insisting on standards that are psychologically and developmentally realistic: it is fruitless to prescribe what we cannot perform. Such a theory also avoids the Charybdis of historical relativism by insisting on a distinction between these standards and the actual performance: it is empty to deflate what is right by identifying it with what is done.

2. Procedural Rationality

Thus, we seek a realistic dynamic model of scientific discovery. We seek to understand the cognitive strategies, the procedures, constitutive of scientific rationality. These strategies are, from one perspective, the procedures that define how humans approach the problem of understanding the world. They define *how* we think about the world. From another perspective, the procedures humans embody constitute assumptions about the structure of the world, or of the part of it to be explained. They define *what* we think about the world.

We take it as axiomatic that rationality is to be understood in terms of problem-solving capacities and skills. Problem solving in general can be understood as a constrained search in a *problem space* that is more or less well defined (cf. Holyoak 1990). In what is by now its canonical form, due to Allen Newell and Herbert Simon (1972), a problem space in a domain includes all possible configurations of the domain, or the number of its possible states. A problem representation, given a problem space, requires four basic components: a *goal state* to be achieved, an *initial state* at which we begin, a set of *operators* defining allowable moves within the state space, and *path constraints* imposing additional limits on what counts as a successful solution. A *solution* is a sequence of operations that leads to the goal state and conforms to the path constraints, and a *problem-solving method* is a procedure for finding a solution for the class of problems at hand. The game of chess provides a useful paradigm for understanding problem solving within a well-defined search space: There are a finite number of possible board positions, each consisting of a permissible arrangement of pieces on the board; these positions define the problem space. Each board position permits only a finite number of alternative moves; these are the path constraints. A solution at any stage would be a series of moves from a given position terminating in checkmate.

This way of understanding problem solving dictates severe practical limits. If a solution requires at least a search through n steps, and if there are M allowable moves at each step, then the number of paths will be M^n. This is literally an astronomical number for even relatively modest values of M and n. If a game of chess involves a total of 60 moves and an average of 30 alternative legal moves at each stage, then the number of paths is 30^{60} or roughly 10^{88} (cf. Newell, Shaw, and Simon 1958). By way of contrast, the universe is something like 15 billion years old. A general brute force solution for chess is obviously out of the question. Some means must be found to limit the search to a computationally tractable number of alternatives. In the face of complex problems, humans typically engage in heuristic search, examining only a small subset of the abstractly possible alternatives. Understanding human problem-solving is, in part at least, understanding the heuristics that guide this search and the way they interact with other variables. To make the point another way, a psychologically realistic model of human problem-solving must respect the limitation imposed by what Simon (1969) calls "bounded rationality": We are organisms limited in memory, attention, and patience. Given these limitations, we limit our search by imposing assumptions about what a solution must look like. These are heuristic assumptions, adopted because of our bounded rationality. Other organisms might impose other assumptions, or other heuristics, but some assumptions are inevitable in the face of complex problems. A psychologically realistic model of human problem-solving must incorporate the heuristic assumptions imposed by human problem-solvers.

Analogously, the task of constructing an explanation for a phenomenon in a given scientific domain is one of finding a sufficient number of variables, the constraints on the values of those variables, and the dynamic laws that are functions on those variables, so that it is possible to predict future states of affairs from descriptions of the universe at an earlier time. A complete state description will pick out a single point in a multidimensional space corresponding to a state of the relevant part of the universe at the time. The number of variables defines the dimensionality of the state space. The dynamic laws are functions from states to subsequent states within this state space. An increase in the number and significance of interactions amounts to an increase in the dimensionality of the problem space needing to be searched. The problem facing a scientific theorist is one of finding laws and variables sufficient to explain what does and does not happen. This is a search in a space of explanatory models, one that will also be subject to a variety of constraints. Some are quite general, such as limitations on human memory. Others are more local, such as limitations on available mathematical models, or simple technological limitations on what data can be gathered. There must therefore be heuristics guiding this search. Understanding scientific problem-solving is, then, a matter of understanding the heuristics that guide the search, both for the relevant variables and the laws.

We also take it as axiomatic that human problem-solving capacities and skills are not domain independent. In some cases, problem-solving methods can be understood as fairly general in application; in others, they are more specialized. Early work in artificial intelligence (AI), including generate-and-test methods, early theorem provers, and Newell and Simon's means-end analysis incorporated methods that could be applied in nearly any domain. They are also weak methods, capable of yielding only relatively poor performance. The moral that the limitations on general search strategies supports is relatively straightforward:

> A system exhibits intelligent understanding and action at a high level of competence primarily because of the specific knowledge that it can bring to bear: the concepts, representations, facts, heuristics, models, and methods of its domain of endeavor. (Feigenbaum 1989, p. 179)

This is by now a generally well accepted moral deriving from work on expert systems and human expertise. Robert Glaser and Michelene Chi explain that, in the light of early work on experts and expert systems,

> it became widely acknowledged that the creation of intelligent programs did not simply require the identification of domain-independent heuristics to guide search through a problem-space; rather, that the search processes must engage a highly organized structure of specific knowledge for problem solving in complex knowledge domains. (Chi, Glaser, and Farr 1988, p. xvi)

We will see that, in the absence of detailed, domain-specific restrictions, the solutions arrived at are rough approximations at best. In Minsky and Papert's terms, a "knowledge-based" strategy is more successful than a "power-based" one (cf. Minsky and Papert 1974, p. 59).[6] An analogous moral extends to human problem-solving. Though some researchers have promoted formal, domain-independent rules (for example, Braine 1978; Rips 1983), it seems reasonably clear that without context-sensitive and domain-specific principles, intelligent problem-solving would be impossible (cf. Cheng and Holyoak 1985, 1989). Rationality should accordingly be understood as requiring the application of specialized cognitive skills and information. Stronger assumptions will limit the search space more, and the more restrictive these assumptions, the more efficient they will be in reaching solutions, if they reach solutions at all; the weaker the assumptions, the more search will be necessary to reach a solution.

Consider the difference between two AI systems, DENDRAL and BACON. DENDRAL is a relatively specialized computer system designed to identify organic molecules from information concerning mass spectrograms and magnetic resonances (Buchanan, Sutherland, and Feigenbaum 1969; Lindsay et al. 1980). After an initial survey identifying selected molecules in a hypothetical sample, DENDRAL generates the set of possible molecular structures on the basis of physical constraints, including such variables as valences and chemical stability. The system then predicts mass spectrograms for these molecular structures and determines a best fit with the data. The generation of possible molecular structures is a procedure allowing for an exhaustive search within the domain defined by the physical constraints. DENDRAL is a model that is highly domain specific, imposing sharp restrictions on the class of solutions, and is very efficient. By contrast, BACON is a program for finding quantitative generalizations and can be applied in a wide variety of fields (see Langley et al. 1987). Given a set of independent and dependent variables, with associated values, BACON looks for correlations that will predict observed values. In its earlier incarnations, BACON would search for constancies and linear relations within the data set and was capable, for example, of inducing Boyle's law from Boyle's own data, Galileo's principle of uniform acceleration, and Ohm's law.

The procedures incorporated in both DENDRAL and BACON are heuristic in motivation and character (for an excellent discussion of heuristics and their role in scientific research programs, see Wimsatt 1980a, 1981, as well as Nickles 1987b), yet they also vary in their power. In some limited cases, in which either the dimensionality of the problem space is low or the problem is simply structured, it is possible to define algorithms that provide an effective and efficient procedure for reaching a solution. Checkers, unlike chess, is a game that lends itself to algorithmic methods: the number of possible board configurations is small in checkers, and the operators are few. In all but the more trivial cases, however, there is no known algorithm that is both effective and efficient for chess or for more complex problems. As a consequence, the most that can be expected is to find procedures that work reasonably well in a limited range of cases. This is accomplished by incorporating assumptions about the domain that limit the number of possibilities to be considered, thereby simplifying the problem-solving task: one limits the search space. In DENDRAL the assumptions about allowable molecular structures limit the search space[7]; in BACON there is a stringent limitation on the form that laws can take, permitting only certain simple, causal relationships to be modeled.[8]

In general, a more domain specific model incorporates assumptions about the domain that are more restrictive and consequently more powerful. For the range of cases defined by such assumptions, the model will be more efficient. On the other hand, a more general model incorporates less restrictive assumptions about the domain, and will therefore be less powerful. In either case, however, the heuristic assumptions imposed are fallible. There will be a range of cases in which a heuristic procedure either will reach no solution at all or will reach an incorrect solution. If we think of the procedure as incorporating a set of assumptions about the task domain, then it will be precisely the cases in which the simplifying assumptions are not met that the procedure will fail. DENDRAL will work only for organic molecules; BACON is limited to laws of a defined set of mathematical forms that, in turn, define the relationships it can see. Thus, not only are heuristic procedures fallible, but their failures are systematic. This dual character of heuristic procedures is an immediate consequence of what makes them both useful and unavoidable; it is the fact that they simplify the problems posed that makes them useful, and it is the simplification that results in systematic failures.

There are parallels to these morals from artificial intelligence in the human case. To take an example that may be relatively familiar, Noam Chomsky has long defended a picture of language acquisition requiring a strong nativist component. The number of possible languages consistent with the sort of data used in learning languages, he claims, is simply too large to be searched effectively without some additional limitations. The nativist component is supposed to specify general features of human languages and thereby limit the search space required to determine the grammar for the language being learned. Chomsky understands language acquisition, in part, in terms of restrictions imposed by this nativist component. At the same time, the nativist component incorporates nontrivial assumptions about the structure of human language, thus imposing assumptions about the structure of the language to be learned. If such nativist procedures are used in learning language, then there exist languages we simply cannot learn precisely because their structures violate the assumptions built into these heuristics.

Our problem in approaching a model for scientific discovery is structurally analogous to work on problem solving. Scientific discovery involves a heuristically guided search for solutions in a complex problem space. However, unlike the games that provide the prototype for problem solving, scientific problems are often ill defined; that is, the constraints defining an adequate solution are not sharply delineated, and even the structure of the problem space itself is unclear (cf. Reitman 1965). As Simon (1973b) points out, an important part of real world problem-solving is, often enough, imposing an appropriate structure on the problem. But, whether well or ill defined, there must be some means for restricting the relevant search space. One way to do this is by imposing some restrictions concerning possible solutions. We may limit the number of variables, assuming only some will make a significant difference. For example, if we are interested in predicting the orbit of a planet, the influences of other planets, for the most part, induce only minor perturbations; however, for other purposes, such as assessing the influence of orbital variations on global climate, their influence is more significant. Alternately, we may impose assumptions about the form of the relevant laws. Thus, in population dynamics it is usually acceptable to assume that the functions will be linear and influences will be additive, although this is not always true. As

Robert May (1974, 1976) initially pointed out, even relatively simple systems will exhibit complex dynamic properties given appropriate values for population size and variables such as birthrate. Limiting the relevant variables and imposing assumptions about the form of relevant laws are procedures for attacking problems. They also describe partial solutions. Applied to scientific problem-solving, this would mean that the heuristic assumptions constitutive of explanatory models would be critical in a developing research program. Without them, it would not be possible to formulate or develop a practical plan of research. The need for heuristic assumptions also raises the prospect that a program of research is misguided, its results simply artifacts of the simplifying assumptions that define the program.

The unification of the procedural and the descriptive will engender philosophical resistance, but it is a natural consequence of a shift away from abstract models of rationality to ones that are sensitive to a demand for realism. If rationality is to be understood in terms of problem-solving capacities and skills, then insofar as human problem-solving capacities and skills are specialized to particular task domains, rationality should be understood as a matter of the application of specialized abilities. This application, in turn, is a matter of incorporating assumptions about the structure of the domain under investigation.

CHAPTER 2

Complex Systems and Mechanistic Explanations

1. Mechanistic Explanation

Our aim is to develop a cognitive model of the dynamics of scientific theorizing that is grounded in actual scientific practice. Our focus is on one kind of explanation, one involved in understanding the behavior of complex systems in biology and psychology. Examples of the complex systems we have in mind are the physiological system in yeast that is responsible for alcoholic fermentation, and the psychological system responsible for memory of spatial locations. As we shall discuss in this section, these explanations, which we refer to as *mechanistic explanations*, propose to account for the behavior of a system in terms of the functions performed by its parts and the interactions between these parts. The heuristics of decomposition and localization are central to our analysis of the development of mechanistic explanations. We shall discuss these heuristics in some detail in this chapter, especially in sections 2 and 3 of this chapter. Parts II and III will be concerned with illustrating their role, and also their development under other influences. These heuristics, or family of heuristics, can be thought of as imposing assumptions about the organization of the systems being explained. We shall examine these assumptions and the kinds of organization of actual systems for which these assumptions are likely to succeed and those on which they will likely fail.

By calling the explanations *mechanistic*, we are highlighting the fact that they treat the systems as producing a certain behavior in a manner analogous to that of machines developed through human technology. A machine is a composite of interrelated parts, each performing its own functions, that are combined in such a way that each contributes to producing a behavior of the system. A mechanistic explanation identifies these parts and their organization, showing how the behavior of the machine is a consequence of the parts and their organization. What counts as mechanistic, though, changes with social context. Scientists will appeal analogically to the principles they know to be operative in artificial contrivances as well as in natural systems that are already adequately understood. The state of technology and natural science at any given time thus plays a significant role in determining the plausibility and limits of mechanistic explanations (cf. Gregory 1981). From the universe of the *Timaeus*, through the Archimedian analogues of Galileo and the clockwork universe of Newton, to the recent focus on servo-mechanisms and computers, the available analogues were important factors in determining which mechanistic models scientists advanced.

The nature and plausibility of mechanistic models is also influenced by characteristics of human thinking, especially by the proclivity of humans to trace operations in a linear or step-by-step fashion. This is especially evident when we consider the forms of organization possible. Many machines are simple, consisting of only a handful of parts that interact minimally or in a linear way. In these machines we can trace and describe the events occurring straightforwardly, relating first what is done by one component, then how this affects the next. Such machines induce little cognitive strain. Some machines, however, are much more complex: one component may affect and be affected by several others, with a cascading effect; or there may be significant feedback from "later" to "earlier" stages. In the latter case, what is functionally dependent becomes unclear. *Inter*action among components becomes critical. Mechanisms of this latter kind are *complex systems*. In the extreme they are *integrated systems*. In such cases, attempting to understand the operation of the entire machine by following the activities in each component in a brute force manner is liable to be futile.

A major part of developing a mechanistic explanation is simply to determine what the components of a system are and what they do. In broad outline, there are two strategies available to analyze and isolate component functions. The first is to isolate components physically within the system and then determine what each does (the goal is to use the knowledge of components to reconstruct how the system as a whole operates). The second strategy is to conjecture how the behavior of the system might be performed by a set of component operations, and then to identify components within the system responsible for the several subtasks. The former is the *analytic* method of Etienne Condillac, which played a role in the development and promotion of both Lavoisier's chemistry and Cuvier's functional anatomy. The latter is the explanatory program of functionalism in contemporary philosophy of mind (see Cummins 1983; Dennett 1978; Lycan 1981a, b), which for contrast we will refer to as a *synthetic* strategy (cf. Posner and McLeod 1982). An analytic strategy constructs from the bottom up; a synthetic strategy projects from the top down.

The analytic strategy is confronted by the fact that smoothly operating systems conceal their component operations. As we will see repeatedly in ensuing chapters, the breakdown of normal functioning often provides better insight into the mechanisms than does normal functioning. In the absence of natural error, failure can be induced. Simon develops the point clearly:

> A bridge, under its usual conditions of service, behaves simply as a relatively smooth level surface on which vehicles can move. Only when it has been overloaded do we learn the physical properties of the materials from which it is built. (1981, p. 17)

Although overtaxing a system induces malfunction, which can be an important clue to the functional properties of component parts, this is a relatively crude method—one as likely to lead to catastrophic breakdown and chaos as to insight. For that reason, researchers prefer to differentiate components and their functions by altering specific activities *within* a system. Such *inhibitory* or *deficit studies* allow us to determine a physical component's contributions to the system by inhibiting its operations and then observing resulting deficits in overall system behavior. The best-known examples of such studies are ablation studies in nineteenth-century physiology (see Harrington 1987), but chemical poisoning studies in biochemistry, and the use of x-rays in the genetic research of Beadle and Tatum, follow the same logic.

Inhibitory studies are problematic for a number of reasons. Perhaps the most serious danger is the temptation to infer, from the fact that a specific experimental manipulation interrupted or inhibited a particular activity of the system, that the part of the system damaged is the component responsible for that activity. As R. L. Gregory pointed out over thirty years ago,

> Although the effects of a particular type of ablation may be specific and repeatable, it does not follow that the causal connection is simple, or even that the region affected would, if we knew more, be regarded as functionally important for the output—such as memory or speech—which is observed to be upset. It could be the case that some important part of the mechanism subserving the behavior is upset by the damage although it is at most indirectly related, and it is just this which makes the discovery of a fault in a complex machine so difficult. (1961, p. 323)

Simplistic uses of functional deficit studies might lead us, for example, to conclude that a resistor in a radio is a hum-suppressor because the radio hums when the resistor is removed (Gregory 1968, p. 99).

Localization based on deficit studies is often erroneous. What is required is some means of figuring out, from the observed deficits, what the component in question positively contributes to the system when it is functioning normally. These are the functions that a good mechanistic explanation will localize in the parts. We will see that an important guide to finding these functions is the simultaneous use of a variety of inhibitory techniques. This can sometimes allow the investigator to determine component functions, but such an account will be complete and compelling only when there is an explanation of how these components interact in effecting the normal operation of the system.

An alternative analytical technique follows an opposite approach, stimulating a physical component and observing the behavioral effects on the whole system. If extra stimulation produces an identifiable surplus, we can sometimes infer that under normal conditions the component was responsible for that which is now generated in excess. We shall generally refer to such investigations as *excitatory studies*. The best-known examples are stimulation studies in neuroscience in which electrical stimulation is applied directly to the cortex. Biochemical studies in which potential metabolic intermediaries are injected exhibit a similar logic. Once again, however, there is a temptation to infer that, because excitation to a physical system enhances or induces a particular effect, the stimulated part is necessarily the seat of the function in question. With both inhibitory and excitatory studies, the natural operation of the system is modified; the experimental techniques may be the source of experimental artifacts and may not be diagnostic of normal operations within the system. They are, however, the only techniques available in some cases. Even lacking a clear conception of the organization of the system, analytical strategies allow the experimentalist to probe the system and its organization.

A synthetic strategy requires some prior hypothesis about the organization and operation of the system. From an initial hypothesis about the underlying mechanisms, one formulates a model of how the system functions. The empirical task involves testing performance projected on the model against the actual behavior of the system. One discipline using model studies is AI, in which researchers propose and implement a set of operations in order to perform an activity that would require intelligence if performed by a human. Such model studies are also used in other domains. In a later section we will discuss biochemical research in which comprehensive proposals were advanced as to the intermediate chemical reactions in an overall physiological process. This theoretical work was sometimes supported by actual development of artificial systems intended to show that a mechanism such as the one being proposed could in fact carry out the needed process.

As plausible as such a model might be, it might also turn out to be radically misconceived. One such case is Justus Liebig's (1842) general account of nutrition. Liebig proposed a model of nutrition which plotted a complete set of metabolic transformations, based only on information about the chemical composition of food and waste products. He was guided by the assumption that because food substances are more complex than waste products, all that an animal does is break down these substances to release the energy stored in them. Using his knowledge of basic chemical reactions, Liebig proposed a bold and brilliant model of animal metabolism. It was also wrong. Though consistent with the data Liebig used, it foundered on physiological data introduced by Bernard, who later showed that animals not only break down complex foodstuff, but also synthesize substances necessary for life (see Bechtel 1982).[1]

The synthetic strategy is traditionally regarded as speculative. Perhaps it is. Speculative stages are nonetheless important to all science, if only in identifying possible mechanisms in contexts in which no workable mechanisms were envisioned. The subsequent testing of these models can and does induce the development and elaboration of more adequate models. However, speculation without empirical constraints is as likely to produce spurious explanations as correct ones.[2] In some cases, one might posit component operations that are, in fact, as complex as the overall operation. We will then have no net explanatory gain. In other cases, one might propose only one among many possible mechanisms and have no warrant for thinking it is the one actually utilized. This is hardly better than having no explanation at all.

The analytic and synthetic strategies are complementary. Inhibitory and excitatory studies can provide empirical data appropriate for evaluating synthetic models. Moreover, synthetic models can provide a theoretical framework in which to interpret information obtained from the empirical studies employing analytical strategies. Even together, though, the strategies hardly provide a fail-safe methodology. The dangers of spurious explanation and premature localization still confront the scientist.

Several factors can make the process of developing mechanistic explanations using techniques such as these extremely challenging. To begin with, as the level and significance of interactions increases, the complexity of the explanatory problems increases as well. The task of constructing an explanation for a given domain might be viewed as one of finding a sufficient number of variables, the constraints on the values of those variables, and the dynamic laws that are functions of those variables. These laws make it possible to use the model to predict future states of affairs from descriptions of an earlier time. The number of variables, once again, defines the dimensionality of the state space or problem space. A complete state description will pick out a single point in a multidimensional space corresponding to a specific value for each variable in the relevant part of the universe at the time. The dynamic laws are functions from states to subsequent states within this state space.

Let us focus just on the question of how complex the state space will be that we need to consider in order to represent the state of the domain. An increase in the number and significance of interactions posited in the constraints and laws requires an increase in the dimensionality of the state space needed to represent the domain. An example may help. Richard

Lewontin (1974) explains that in the case of genetics we can think of the problem as one of predicting the genetic composition of a population from its composition at an earlier time. As we consider greater numbers of loci and alleles, the number of variables needing to be considered increases and, hence, so does the dimensionality of the state space. What is pertinent to our purposes is how interaction affects the number of variables, and consequently the number of dimensions in the state space. If we assume that genes at different loci segregate independently, then the predictive problem reduces to one of finding solutions to independent problems for each locus, and accordingly the proper unit of analysis will be allelic frequencies. In that case, the dimensionality of the space needed to represent the state of the genome is a linear function of the number of alleles and the number of loci. With *n* loci and *a* alleles at each locus, the dimensionality of the problem space approximates the product of the two. If, on the other hand, we assume that there is significant linkage between genes at different loci, then, Lewontin urges, the proper unit of analysis will be gametic frequencies. As the number of loci increases, the number of possible gametes increases exponentially: in the case of two alleles at each locus, with two loci, there will be four gametic classes; with three loci, there will be eight gametic classes; and with ten, there will be over a thousand. The expected dimensionality for the state space will be one less that the number of allelic or gametic classes. This means that the anticipated dimensionality of the state space needed to represent the state of the genome will also increase linearly with allelic classes as the unit, and exponentially when using gametic frequencies. The results are detailed in Table 2.1.

Explanatory demands are further aggravated by the relative stability of the systems involved. This is part of what helps to conceal the contributions of the individual components when we examine a normally operating system. As we noted previously, if the goal is to reveal what the parts contribute to the operation of the whole system, we generally must find techniques to perturb the system. When investigating the behavior of *self-organizing* systems, the theorist must contend with the fact that these systems will maintain or determine an equilibrium even in the face of considerable perturbations, both internal and external. Consequently, establishing what the parts contribute will be difficult. In the limiting case, self-organizing systems will be homeostatic—that is, they will maintain a predetermined equilibrium state despite perturbations. The genetic system, in which endonucleases repair damage to nuclear DNA, is one example of such a homeostatic system. In more complicated cases, the systems are better thought of as *self-regulating*—that is, they will modulate activity at varying levels depending on other influences. For example, coenzymes in cell metabolism adapt the breakdown of foodstuff to the work performed by the cell.

n	$a = 2$		$a = 5$		$a = 10$	
1	1	**1**	4	**4**	9	**10^{+1}**
2	2	**3**	8	**24**	18	**10^{2}**
3	3	**7**	12	**124**	27	**10^{3}**
4	4	**15**	16	**624**	36	**10^{4}**
5	5	**31**	20	**3,124**	45	**10^{5}**
6	6	**63**	24	**15,624**	56	**10^{6}**

Table 2.1. Increase in the Dimensionality of a Problem in Genetics as a Function of Linkage. With *n* loci and *a* alleles at each locus, the dimensionality of the problem depends critically on linkage. Assuming independent assortment (indicated in plain text), the proper unit of analysis will be allelic frequencies. The dimensionality of the state space increases as a linear function of *n*. With linkage effects (indicated in bold text), the proper unit of analysis is gametic frequencies. The dimensionality then increases exponentially. (Based on Lewontin 1974.)

When a system's behavior is relatively constant despite variations external to it, we can safely ignore the environment and focus only on the internal mechanisms to specify the parts, their interactions, and their contributions to the behavior of the system. But when the system adapts to the environment, as homeostatic and self-regulating systems do, we cannot simplify the task in this way. The sensitivity to environmental changes means that the parts operate differently under altered conditions and so further conceal from view how they behave when the system is operating normally. In general, interaction among the various components makes it difficult to isolate independent contributions from their coordinated output.

2. Decomposition and Localization

We now turn to the heuristic strategies of decomposition and localization. As we will see in the chapters to follow, these strategies have been used by a wide variety of researchers in a wide variety of disciplines, from nineteenth-century brain science and early twentieth-century investigations into chromosome structure to more recent work on language and cognition.

Decomposition allows the subdivision of the explanatory task so that the task becomes manageable and the system intelligible. Decomposition assumes that one activity of a whole system is the product of a set of subordinate functions performed in the system. It assumes that there are but a small number of such functions that together result in the behavior we are studying, and that they are minimally interactive. We start with the assumption that interaction can be handled additively or perhaps linearly.

Whether these assumptions are realistic or not is an open question; indeed, at the outset we often simply do not know. The extent to which the assumption of decomposability is realistic can be decided only a posteriori, by seeing how closely we can approximate system behavior by assuming it. We may be led to erroneous explanations, but it may be the only way to begin the task of explaining and understanding complex systems. The failure of decomposition is often more enlightening than is its success: it leads to the discovery of additional important influences on behavior.

Localization is the identification of the different activities proposed in a task decomposition with the behavior or capacities of specific components. In some cases we may be able to identify (through fairly direct means) the physical parts of the system in which we can localize different component functions. In other cases we may have to rely on various functional tools for determining that there are such parts, without being able to identify them; for example, we may be able selectively to inhibit their operation and observe the consequences on behavior. We need not assume that a single part in this sense is a spatially contiguous unit; in fact, we know that in many cases it is not. A functional unit may be distributed spatially within the system. Localization does entail a realistic commitment to the functions isolated in the task decomposition and the use of appropriate techniques to show that *something* is performing each of these functions.

In one extreme form, decomposition and localization assume that a single component within the system is responsible for some range of phenomena exhibited by the system. For example, it is assumed that the posterior cerebral lobes are responsible for vision, that the cell nucleus is responsible for genetic control, and that there is a specialized enzyme responsible for catalyzing a given chemical reaction within the cell. This, the simplest assumption, often guides the first explanatory models, even if it seldom survives. We refer to it as *simple* or *direct localization*, being simple both in focusing on single components and in imposing the fewest constraints. The simplest case, however, is often far too simple (cf. Ch. 4). The behavior to be explained may be at best the product of several independent components rather than of one. No one component can be assigned sole responsibility. It also may be necessary to assume there is some interaction and some differentiation of function. Lacking simple localization, the alternative is to localize a set of component functions and assume linear interaction will explain the behavior of the system. We refer to this as *complex* or *indirect localization*, having not only a complex organization, but complex constraints on the problem.

Pursuing decomposition and localization is to impose an assumption about the nature of the system whose activities one is trying to explain: it is assuming that it is *decomposable*. A decomposable system is modular in character, with each component operating primarily according to its own intrinsically determined principles. Thus, each component is dependent at most upon inputs from other components, influences other components only by its outputs, and has a specific, intrinsic function. The notion of decomposability stems from Simon and constitutes the descriptive counterpart of localization. Localization presupposes that we are confronted with a modular organization such that the components of the system can be subjected to separate study and investigation; it requires that the components have discrete intrinsic functions intelligible in isolation, even if such functions do not independently replicate those of the system as a whole.

An extreme form of a decomposable system is an *aggregative system*,[3] which is a species of *simply decomposable* system. System behavior in such cases is a linear or aggregative function of component behavior (cf. Levins 1970, p. 76). Whatever organization is present is not a significant determinant of the relevant systemic properties. In an important paper, "Forms of Aggregativity," Wimsatt lays down four conditions of aggregativity (1986a, pp. 260–68):

1. Intersubstitutability of parts;
2. Qualitative similarity with a change in the number of parts;
3. Stability under reaggregation of parts; and
4. Minimal interactions among parts.

We emphasize the last condition, primarily because the focus of our investigation is the discovery of organizational properties that fix the interaction of the parts and determine their significance for system behavior. However, we do not see a useful way to elaborate on this condition independently of the first three, and we think that—with a suitable decomposition—when this last condition is satisfied the other three generally will be satisfied as well.[4] In offering the final condition, Wimsatt gives us this statement, with some reservations: "There are no cooperative or inhibitory interactions among parts of the system" (ibid., p. 269). As an example, lateral inhibition in the retina leads to a lower activation among neurons than would be predicted if we attended only to their stimulation level; this is because an elevated activation level for one neuron decreases the activation level of adjacent neurons.

Few interesting dynamic systems are strictly aggregative. For example, fluid flow or the movement of a herd *approximates* aggregative motion, though even these *only* approximate aggregative systems and some of the most interesting work concerns why this is so. When the relevant systemic properties are at least partially determined by the organization of the system, we no longer have aggregativity. Or, more realistically, to the extent that organization determines systemic behavior, the system is nonaggregative. In the simplest departures from aggregativity, we may still maintain intersubstitutability; however, when this also fails, we have what we call *composite systems*. There are two species of composite systems, which differ in terms of the role played by systemic organization: In *component systems*, the behavior of the parts is intrinsically determined. In these cases it is feasible to determine component properties in isolation from other components, despite the fact that they interact. The organization of the system is critical for the functioning of the system as a whole, but provides only secondary constraints on the functioning of constituents. In *integrated systems*, systemic organization is significantly involved in determining constituent functions. There may be, for example, mutual correction among subsystems, or feedback relations that are integral to constituent functioning. Thus, as we will see, although work on cell metabolism by, for example, Neuberg and Thunberg in the early decades of this century treated metabolic processes as linear and sequential, the discovery of coenzymes and their function has made it clear that linear models are oversimplified in the extreme (see Bechtel 1986a, and Chapter 7). Systemic organization then provides primary constraints on constituent functioning, and constituent functioning is no longer intrinsically determined. Some of the most interesting and bewildering problems arise with such systems. Richard Levins comments:

> This is a [type of] system in which the component subsystems have evolved together, and are not even obviously separable; in which case it may be conceptually difficult to decide what are the really relevant component subsystems. Thus, for example, we might consider that a simpler multicellular organism is composed of cells, and yet the cells may be more profitably regarded, under other circumstances, as simply spatial subdivisions, partly isolated, of an organism. (1970, p. 77)

Dependence of components on each other is frequently mutual and may wholly blur any distinction among them. Thus, mitochondria were once independent organisms—though they are now clearly but parts of a cell, integrated into cell metabolism (Margulis 1970), and we cannot now understand how they function if we neglect their incorporation in, and integration into, the complex activities of the cell.

Composite and integrated component systems correspond to two types of organization in Simon's scheme (component systems are Simon's *nearly decomposable* systems). To the extent that components perform independent functions and send their outputs to other parts, we have *strict* or *simple decomposability*. *Near decomposability* imposes less stringent limits, as Simon explains:

> (1) In a nearly decomposable system, the short-run behavior of each of the component subsystems is approximately independent of the short-run behavior of

> the other components; (2) in the long run the behavior of any one of the components depends in only an aggregate way on the behavior of the other components. (1969, p. 210)

For near decomposability, individual components must be controlled by intrinsic factors, in the manner of composite systems. As components become less governed by intrinsic factors, we enter the domain of integrated composite systems, which are *minimally decomposable.*

A system will be nearly decomposable to the extent that the causal interactions *within* subsystems are more important in determining component properties than are the causal interactions *between* subsystems (cf. Simon 1969, p. 209). Wimsatt (1972, p. 72) suggests that we characterize such systems in terms of a parameter that is a "measure of the relative magnitudes of intra- and inter-systemic interactions for these subsystems." As we noted above, the heuristics of decomposition and localization assume a degree of decomposability. At a minimum they assume that the system is at least nearly decomposable. A parameter of the sort Wimsatt offers, then, would be an estimate of the likely error in predictions based on models developed using decomposition and localization (for further discussion, see Richardson 1982).

Whether decomposition and localization will succeed or fail in a given case, these heuristics are important because they provide us with a tractable strategy for attacking the explanatory problems complex systems present. Recent research in the psychology of judgment indicates that humans have great difficulty comprehending cases with more than a few interacting variables. Humans *cannot* use information involving large numbers of components or complex interactions of components, and even when the problem tasks are computationally tractable, human beings *do not* approach them in this way. Complex systems are computationally as well as psychologically unmanageable for humans.

3. Hierarchy and Organization

We need now to explore possible reasons for thinking that the sorts of systems encountered in biology or psychology are likely to be decomposable or nearly decomposable—and hence amenable to mechanistic explanations developed through decomposition and localization—or only minimally decomposable or not decomposable at all—and hence not amenable to explanation using these heuristics. To explore this issue it is useful first to recognize that in talking about components and whole systems we are construing nature as incorporating a hierarchy of levels. When parts interact with each other, we can view this as a *horizontal* process; that is, there is interaction between units at one level. When we focus on how components combine into larger units, which in turn may interact with other larger units, we are addressing a *vertical* question about the relations between levels. If the degree of interaction among components when they come to form wholes does not obliterate the components as autonomous entities, the result is a *decomposable part-whole hierarchy*.

A hierarchical organization also facilitates tractability by providing for theoretical economy. In explaining the behavior of the system we can gain independent characterizations of each component, ignoring both the contributions of other components at the same level as well as the influences operative at higher or lower levels. For example, programing in higher-level languages allows us to bypass the means by which the commands are executed. If the system is decomposable, there will be relatively little information lost in such a representation: "In studying the interaction of two large molecules, generally we do not need to consider in detail the interactions of nuclei of the atoms belonging to the one molecule with the nuclei of the atoms belonging to the other" (Simon 1969, p. 218).

Moreover, there is considerable evidence that information is held in human memory within organized structures and that the very intelligibility of a domain may depend on its being represented as a decomposable hierarchy. In work by Simon and his collaborators, it has been shown that experts in various disciplines differ from novices not just in the amount they know, but in the way their knowledge is organized. Expert chess players, for example, are readily able to reconstruct board positions of games that they have observed for only a few seconds, whereas novices are able to locate only a few pieces. This difference disappears, however, when the board positions do not make strategic sense. Chase and Simon (1973a, b) contend that the differences are due to the fact that experts recognize and remember patterns among pieces and treat these patterns as units (see also Gilmartin and Simon 1973). Analogously, in tasks evaluating the efficiency of recall and recovery, memory is facilitated by texts with specific forms of organization and inhibited by others. Free recall will even impose organization when none is present in the text.

Simon holds that hierarchical organization is also a general phenomenon in nature. He argues that hierarchies arise because the forces governing interactions between objects typically do not form an equally distributed continuum. The strongest forces govern interactions at the lowest level and give rise to reasonably stable units at a middle level. In many cases these lower-level forces may not determine which among a variety of complexes at the higher level will be realized. They constrain but do not determine the results. Weaker forces then come to play in determining the relationships between these middle-level entities.

As an illustration, Simon considers chemical forces. The forces responsible for atomic structure are stronger than those that determine the composition of the molecules made of them. Similarly, the forces determining the structure of macromolecules are weaker than those that determine their composition.

> Thus, protons and neutrons of the atomic nucleus interact strongly through the pion fields, which dispose of energies of some 140 million electron volts each. The covalent bonds that hold molecules together, on the other hand, involve energies only on the order of 5 electron volts. And the bonds that account for the tertiary structure of large macromolecules, hence for their biological activity, involve energies another order of magnitude smaller—around one-half of an electron volt. It is precisely this sharp gradation in bond strengths at successive levels that causes the system to appear hierarchic and to behave so. (Simon 1973a, p. 9)

An equilibrium between the forces at the lowest level defines a set of stable systems. The forces at the higher level then determine the relationship between these units and their combination into other units. In biochemical processes it has been clear since the work of Linus Pauling and Max Delbrück (1940) that quantum mechanical forces are insufficient to explain reactions among complex molecules in the cell; much weaker forces are required for the intermolecular interactions. Hydrogen bonds are especially important, both in antibody formation and in the classic work on the double helix by Watson and Crick (1953). The resulting structures again form into stable units, with their interrelations defined by still weaker forces.

The argument so far for nature comprising decomposable hierarchies assumes that the strengths of the forces for binding components into structures are not continuously distributed. This may or may not be true, but Simon offers other, more general, arguments. For example, he appeals to evolutionary considerations, arguing that complex systems are more likely to evolve if they are hierarchical and decomposable. He assumes that the lower-level forces insure that components can arise independently, existing as stable units in their own right. All that would then be necessary is the formation of stable combinations that would meet the demands at the higher level. Given complex macromolecules, living cells would be combinations formed from them, dependent only on the operation of higher-level forces. Selection could then serve to fine-tune the system without large-scale disruption. To quote once again from Simon,

> The loose horizontal coupling of the components of hierarchic systems has great importance for evolutionary processes just as the loose vertical coupling does. The loose vertical coupling permits the stable subassemblies to be treated as simple givens, whose dynamic behavior is irrelevant to assembling the larger structures, only their equilibrium properties affecting system behavior at the higher levels.

> The loose horizontal coupling permits each subassembly to operate dynamically in independence of the detail of others; only the inputs it requires and the outputs it produces are relevant for the larger aspects of system behavior. (1973a, p. 16)

To illustrate the principle, Simon (1969) tells the following tale of two watchmakers: Each makes fine watches with 1000 parts of diverse sorts. One of the two, Tempus, uses an hierarchical design with component parts: each watch has ten components, and each component has ten components, etc. The other, Hora, uses a horizontal design. All the parts must be in place before the watch will stay together. Each of the watchmakers is interrupted periodically to take orders for additional watches. Hora's work suffers dramatically, because every interruption results in a loss of all the work on the current watch. Tempus' work suffers too, though not so dramatically, because all that is lost is the work on the current subassembly. The moral is a general one: complex structures are more efficiently constructed if they are composed of stable subassemblies. Simon shows that additional levels of intermediate structure will further increase stability; consequently, the time required to assemble a system with any given number of units is inversely proportional to the number of intervening levels. He then goes on to apply the same principle to the evolution of biological units: decomposability increases the evolutionary rate (ibid., pp. 200ff.). As a result, complex systems arise more readily when they consist of stable subsystems.

As we noted in Chapter 1, the assumption that a system is nearly decomposable and hierarchical is not just motivated by theoretical arguments; human cognitive strategies make such an assumption natural. This does not mean, however, that it is realistic. There are in fact reasons to suspect that many natural systems are not decomposable even though they are hierarchical. Simon concentrates on the division of systems into component parts. In minimizing interconnections between these parts, and treating them as autonomous, Simon sidesteps discussion of what binds parts together, making them *parts* of a complex system. If nothing imposes systemic structure, we have an aggregative system. Systemic behavior is an additive function of component behavior, at least if we ignore threshold effects. We have a hierarchy in name only. With composite systems, interaction between units is critical; indeed, it is constitutive of the higher-level units. Interaction is what makes composites useful for explanatory purposes; however, interaction also compromises the autonomy of components.

The mode of organization is important. Grobstein (1973) distinguishes between *facultative* and *obligate* organization, in a transparent analogy to social symbioses (cf. Richardson 1982). A facultative organization allows members to disperse and recombine. Individuals can function as part of a more complex system, but are also capable of independent activity; for example, baboons will often forage in groups, but they are also able to forage independently. Facultative organization is thus nearly decomposable. An obligate organization, by contrast, is one in which interdependence has significantly compromised the capacity for independent activity, and the system is thus only minimally decomposable. As Grobstein says, the properties of higher-level structures "are in some sense immanent in the properties of the components, [though] . . . such properties tend to be lost if components of a set are dispersed or if a set is dissociated from its context or superset" (1973, p. 45). Some flowering plants are wholly dependent on birds or insects as vectors for pollination; and some of these birds and insects, in turn, are specialized, feeding on only one type, or a few types, of flowering plants. The latter form of organization is particularly important, as Grobstein recognizes:

> Their components are very different in properties when in isolation or in the collective, and the collective, once formed, is not reversible. This is the case with most higher organisms. . . . A complex multicellular organism represents an extreme case in which very special conditions are required to maintain individual cells or individual organs outside of the collective relationship. (Ibid., p. 34)

Not only are the relationships among constituents of the system important for explaining the system's operation, but the constituents themsleves have no independent, isolable function.

Simon's theoretical arguments supporting the ubiquity of nearly decomposable systems with facultative organization rest on evolutionary considerations. Such considerations, though, will not support the conclusion that complex systems are generally decomposable. Levins (1970) and Wimsatt (1972) point out that divergence and coadaptation will decrease the decomposability of a system with time: once aggregated, components can diverge and specialize in functions while maintaining stability. Evolvability does not insure stability; thus, while decomposable hierarchies may be more likely to evolve, the considerations advanced do not necessarily favor maintaining decomposability once they are formed. It will remain an open question whether complex natural systems are necessarily decomposable hierarchies.

4. Conclusion: Failure of Localization

Whether natural systems are hierarchically structured will influence how successful decomposition and localization will be as heuristic strategies. In this chapter we have described decomposition and localization as heuristics for developing mechanistic explanations of complex systems and have examined the assumptions these heuristics make about the nature of the system we need to explain. While there are considerations favoring the occurrence of decomposable hierarchies, there are also considerations pointing toward only minimally decomposable, integrated systems. Thus, there are clearly risks in assuming complex natural systems are hierarchical and decomposable. There are always some risks that stem not from the assumption of decomposable hierarchies, but from specific errors in developing the decomposition and localization. In a case of false localization, a complex system may manifest a component organization, but not the specific component analysis attributed to it. In this case the way the system operates is misrepresented. This may be because there is an alternative component organization at the same level, or because we have adopted the wrong level of analysis altogether. We will consider cases subsequently in which the initial analysis is misguided in these ways.

More radical errors are also possible. The separation of systems into isolated components, with the attendant minimization of interactive importance, may blind us to critical factors governing system behavior; in particular, it may blind us to the importance of systemic interaction. We may not have a decomposable system at all, or we may have one that is only minimally decomposable. Simon acknowledges the risk saying, "If there are important systems in the world that are complex without being hierarchic, they may to a considerable extent escape our observation and understanding" (1969, p. 219). The risk, if realized, should be felt in failures of explanation. If the failures are more limited, we are only limited in our explanations. We will also examine failures of a more radical sort in which, though research began with the assumption of near decomposability, high degrees of organization were subsequently recognized and the explanatory approach had to be adapted to accommodate this organization. Though decomposability may be a natural and fruitful starting point, it may be no more than that.

PART II

Emerging Mechanisms

It has been found in science that when a sub-universe of discourse can be dissociated from a larger universe and a means of studying behavior found which is but slightly affected by uncontrollable factors, the results usually have a high value in prediction.

—E. M. East 1934

Every organized being forms a whole, a unique and closed system, whose parts mutually correspond and concur to the same definite action by a reciprocal reaction. None of its parts can change without the others also changing; and consequently each of them taken separately, indicates and determines all the others.

—G. Cuvier 1812