Chapter 1

Computational Approaches to Scientific Discovery

Jeff Shrager
Pat Langley

1. Introduction

Science is perhaps the most complex of intellectual activities, and its study has traditionally been the realm of historians and philosophers. However, recent advances in cognitive science—particularly in artificial intelligence and cognitive psychology—have provided new approaches and fresh insights into the nature of science. Whereas early work in the philosophical tradition emphasized the *evaluation* of laws and theories (e.g., Popper, 1965), recent research in the paradigm of cognitive science has emphasized scientific *discovery*, including the activities of theory formation, law induction, and experimentation. Moreover, the early philosophical approaches focused on the *structure* of scientific knowledge, whereas recent work has focused on the *process* of scientific thought and on describing these activities in *computational* terms. The aim of this chapter is to provide an overview of this computational research on scientific discovery.

Three basic developments have led to progress in this area during the past decade. First, cognitive psychology has made significant advances in its understanding of complex human behavior, which have encouraged psychologists to study domains such as scientific reasoning (e.g., Gholson, Shadish, Neimeyer, & Houts, 1989; Mynatt, Doherty, & Tweney, 1978, this volume; Shrager & Klahr, 1986). Second, the field of artificial intelligence has evolved into a mature discipline and has explored

and learning (e.g., Weld & de Kleer, 1990). Finally, many philosophers of science have adopted a historical and psychological perspective on science, focusing less on normative and structural theories and more on how discoveries actually take place (e.g., Darden, this volume; Kuhn, 1962; Lakatos, 1970; Thagard & Nowak, this volume).

These advances have supplied the data and techniques needed to construct detailed computational models of the acquisition of knowledge in scientific domains. Research goals and methods differ, with some researchers giving detailed accounts of historical discoveries, others studying subjects' behavior in simulated scientific settings, and still others—caring less for historical or psychological adequacy—proposing algorithms with desirable computational properties. Taken together, these different emphases provide a multifaceted view of scientific discovery, giving a broader and deeper understanding than was possible even a few years ago.

We begin our survey of computational models of discovery by identifying some components of scientific behavior and proposing an associated vocabulary. We then review recent progress in computational approaches to discovery, using our framework to describe developments during the past five years. Finally, we consider some open problems in scientific discovery that do not fall within the framework and that have not been modeled in existing systems. We argue that these issues should receive significant attention in future research.

2. Components of Scientific Behavior

In order to discuss computational theories of scientific behavior, we need a vocabulary with which to describe their components. In English, such terms as *discovery* and *theory formation* describe the diverse and complex behavior of a scientist at work, but in a vague and ill-defined manner.[1] One advantage of computational approaches is that they force the researcher to provide precise specifications of data structures and algorithms. Unfortunately, the goal of implementation often leads one to adopt narrow definitions of concepts that potentially have a much wider scope.

[1] We will assume that the scientist is working *alone* in a given domain and that he or she has instruments available to manipulate and observe the domain. Later,

Following the tradition in artificial intelligence, we divide scientific behavior into *knowledge structures* and the *processes* or activities that transform them. Although narrow definitions are necessary to produce computational models, they are not required in a survey. Therefore, we will restrict ourselves to definitions of knowledge structures that are independent of particular representations, and to definitions of activities that focus on input/output relations rather than on specific methods. Even at this level, clear definitions are difficult to provide, and the reader should treat the statements that follow as tentative formulations. In addition, the list of components is clearly incomplete, being limited to aspects that have been addressed in existing models.

In the following discussion, we assume that the scientist is working in some particular *field* and more specifically on some problem in a particular *domain* within that field. For instance, the domain of neutrino interactions lies within the field of nuclear physics. We further assume that the scientist is operating in a laboratory or in some other relatively controlled *setting* (as opposed to field work), and we refer to particular arrangements of the setting, such as a specific experimental arrangement, as a *situation*. All of this together will be called the scientific *environment*.

2.1 Scientific Knowledge Structures

Before we can talk about activities, we must identify the knowledge structures that are inspected and manipulated. Together with the physical setting, these components constitute the raw materials and the products of science. In a given computational model, a number of these structures are cast in some specific representational framework, but in our quest for generality we will avoid commitment to particular representational assumptions. We describe the basic knowledge structures below.

Observations (or *data*) represent recordings of the environment made by sensors or measuring instruments. For instance, in his studies of heat, Joseph Black (1728–1799) recorded the temperatures of objects before and after he heated them. Each of these recordings was an observation.

Taxonomies define or describe concepts for a domain, along with specialization relations among them. One example is the taxonomy for

and so forth. Another is the grouping of chemical substances into acids, bases, and salts, and the subdivision of bases into alkalis and metals. Taxonomies specify the concepts used in stating laws and theories, and in giving units to observations.

Laws are statements that summarize relations among observed variables, objects, or events. For example, Black's heat law states that if one mixes two substances, the temperature of one substance increases and the temperature of the other decreases until they reach equilibrium. It also describes a precise numeric relation among the initial and final temperatures. The first statement is qualitative in form, whereas the latter is quantitative. Some laws may be quite general, whereas others may be very specific, potentially composed entirely of constants or ground terms.

Theories represent hypotheses about the structures or processes in the environment. They differ from laws in making reference to unobservable objects or mechanisms. For instance, the caloric theory stated that all material objects contained a substance called *caloric* and that heating involved a transfer of caloric to the heated object. A theory is stated in terms of concepts from the taxonomy.

Background knowledge is a set of beliefs or knowledge about the environment aside from those that are specifically under study. Such knowledge differs from theories or laws, in that the scientist holds background knowledge with relative certainty rather than as the subject of active evaluation. Statements that begin as theories or laws may eventually come to act as background knowledge. For instance, Black probably assumed that placing a flame under an object would increase its temperature.

Models are descriptions of the environmental conditions, both overt and hidden, for an experimental or observational setting. Thus, a model is required to indicate the manner in which a law or theory applies to a particular situation. For instance, one might attempt to understand a particular physical situation in terms of frictionless pulleys connected by massless strings, thus enabling the application of simple Newtonian mechanical theories.

Explanations are narratives that connect a theory to a law by a chain of inferences appropriate to the field. In such cases, we say that the the-observation that objects of different temperature move toward equilibrium when placed in contact. In some disciplines, inference chains must be deductive or mathematical, but many fields sanction other forms of explanation.

Predictions represent expectations about the behavior of the environment under specific conditions. One prediction that follows from the caloric theory is that a heat source will eventually stop transferring heat since ultimately the source will run out of caloric. For instance, if rubbing two objects together adds heat to the surrounding air, eventually this heating effect will halt. *Postdictions* are analogous to predictions, except that the scientist generates them after making the observations he or she intends the postdictions to explain. Successful predictions and postdictions lend support to the theory or law that produced them.

Anomalies describe laws that cannot be explained by a theory, or observations that cannot be predicted by a law. For instance, suppose one finds that the heating effect continues no matter how long one rubs two objects together. This finding is an anomaly with respect to the caloric theory since that theory leads to no laws that accord with the observation.

Although each of the above concepts plays an important role in scientific thought and action, many developers of specific discovery systems have collapsed some of them and left others implicit. For instance, rather than being stored as a separate structure, a theory might be implemented as an active subset of the background knowledge. Similarly, predictions need not be explicitly represented for one to obtain observations that violate the theory. To our knowledge, no existing discovery system explicitly incorporates all of these concepts.

Before proceeding to the activities of the scientist, we should note some concepts that we have explicitly left out of the foregoing analysis. These include *hypotheses*, *explorations*, *instruments*, and *representations*, as well as many others. Although these are important aspects of science, we believe that the concepts described above provide a satisfactory basis for a concrete discussion of scientific behavior.

2.2 Scientific Activities

Knowledge structures alone cannot provide a complete account of science. Activities that modify these structures and that change the setting

under study are essential to the production of scientific knowledge. In this subsection, we propose a set of activities that describe the classical view of science, though we will broaden this set later in the chapter. Many philosophers have either explicitly or implicitly proposed categories of scientific activities (e.g., Feyerabend, 1975; Hacking, 1983; Lakatos, 1976; Popper, 1965; Suppe, 1977), but only a few computationalists have explicitly addressed this issue (e.g., Falkenhainer & Rajamoney, 1988).

We have attempted to describe activities that lie at approximately the same level. We have also aimed for functional definitions that are specified in terms of the knowledge structures each activity inspects and affects. Any given computational system will use a specific method to implement such activities, but we have intentionally avoided giving particular methods in our definitions. We describe the basic scientific activities below.

The *observation* process inspects the environmental setting by training an instrument, sometimes simply the agent's senses, on that setting. The result is a concrete description of the setting, expressed in terms from the agent's taxonomy and guided by the model of the setting. Since one can observe many things in any given situation, the observer must select some aspects to record and some to ignore. For example, Joseph Black observed a setting in which two fluids were brought into contact. Using a thermometer and a clock, he measured the temperature of each fluid at successive points in time. From this activity, he obtained data providing a set of concrete descriptions of the setting.

Taxonomy formation (and revision) involves the organization of observations into classes and subclasses, along with the definition of those classes. This process may operate on, or take into account, an existing taxonomy or background knowledge. For instance, early chemists organized certain chemicals into the classes of acids, alkalis, and salts to summarize regularities in their taste and behavior. As time went on, they refined this taxonomy and modified the definitions of each class. Another example of changing taxonomies involves the distinction between heat and temperature, which scientists had initially confounded (Carey & Wiser, 1983).

Inductive law formation (and revision) involves the generation of empirical laws that cover observed data. The laws are stated using terms from the agent's taxonomy and are constrained by a model of the setting and possibly by the scientist's background knowledge. In some cases, the scientist may generate an entirely new law; in others, an existing law may be modified or extended. For instance, Black arrived at his law of specific heat to summarize the temperature changes he observed in his heat experiments. Similarly, based on systematic experiments with the pressure and volume of gases in containers, Robert Boyle (1627–1691) induced a law that related these two variables.

Theory formation (and revision) stands in the same relation to empirical laws as does law formation to data. Given one or more laws, this activity generates a theory from which one can derive the laws for a given model by explanation. The theory is stated using terms from the domain's taxonomy and may be influenced by its background knowledge. Thus, a theory interconnects a set of laws into a unified theoretical account. For example, Boyle's law describes the inverse relation between the pressure and volume of a gas, whereas Charles' law states the direct relation of its temperature and pressure. The kinetic theory of gases provides an elegant explanation for both laws in terms of Newtonian interactions among molecules. Theory revision takes into account an anomalous phenomenon or law that cannot be explained by an existing theory. The revised theory should explain the anomalous phenomenon while maintaining the ability to cover existing laws, although this is often not possible.

Deductive law formation produces laws by a second route, starting with a theory and using an explanatory framework to deduce both a law and an explanation of how that law derives from the theory. Recall that laws can be composed entirely of ground terms, so this process can create very specific laws that lend themselves to prediction and thus aid in theory evaluation. For instance, Einstein's theory of general relativity led to an inferred law about the orbit of Mercury. However, not all such derived laws will be testable.

The *explanation* process connects a theory to a law by a narrative whose general form is given by the field's explanatory framework. In the context of evaluation (described below), if such a narrative can be produced, support may be lent to the theory or law from which the prediction arose. If no such narrative can be produced—that is, if explanation fails—then an anomaly results. The explanation process can also aid

to known laws in the domain.[2] Explanation differs from deductive law formation, in that explanation attempts to account for a law that is already known.

The *prediction* process takes a law and a model of the setting, and produces a prediction about what will be observed. This often involves the results of intentional experimental manipulation, but it can also occur in observational domains. For example, one can use the ideal gas law to predict that, upon compressing a cylinder of gas, its temperature will rise. One can also use Kepler's laws of planetary motion to predict that an eclipse will occur at a certain time. The analogous process of *postdiction* takes place in cases where the scientist must account for an existing observation. Prediction and postdiction stand in the same relation to each other as deductive law formation and explanation.

Experimental design generates models of settings in which observations are to be made. Typically, selected aspects of the model (the independent variables) are systematically varied to determine their effect on other aspects (the dependent variables). This design process may take existing laws or theories into account, or it may be more exploratory in nature. Thus, Black decided to systematically vary the substances used in his experiments to determine their effects on rates of temperature change. If competing theories are considered in experimental design, they generally make different predictions.

The *manipulation* process constructs a physical setting that corresponds (to whatever extent possible) to a desired model. Thus, the scientist manipulates the environment in order to implement a given experimental design. For instance, Black instantiated his experimental design for studying temperature phenomena by physically heating various substances.

Evaluation, comparing a prediction with observations, generally follows experimental design and observation. Since predictions can vary in their level of detail, evaluation may vary in what is accepted. This produces either a successful *postdiction* or an *anomaly*, which may serve to stimulate further theory or law formation or revision. For instance, the

[2] A subtlety of the present definition arises from the fact that we have defined the explanatory process to operate on laws, whereas one may sometimes want to explain precise observations as well. However, recall that laws can vary in their level of generality, so that one can easily transform observations into very specific

anomalous behavior of rubbed objects (as described above) shed doubt on the caloric theory.

For the sake of simplicity, we have omitted a number of important activities from the above framework. These include: the process of *accepting* a tentatively held theory, thus adding it to one's background knowledge; the process of *scientific revolution*, in which one revises an entire theoretical framework; *model formation and revision*, in which one generates or revises a model that connects a theory and its laws to an experimental setting; and activities attending the important *social* and *embodied* aspects of scientific activity, such as communication, note taking, perception, and the construction of measurement instruments. In Section 4, we will return to the last of these topics in an effort to expand the traditional view of scientific behavior.

In any particular research endeavor, many of the activities described, as well as those that we have omitted, will be composed into greater units at various levels, ranging from daily actions to weekly plans to research programmes that cover months or years. Specific computational models implement certain combinations of these activities. In surveying the past decade of research on computational models of discovery, we will discuss the particular knowledge structures and activities that researchers have implemented.

3. Recent Research on Machine Discovery

We have chosen to divide research on scientific discovery into two broad periods. The first interval, during which cognitive scientists developed the first computational models of the discovery process, extends from the late 1970s through 1984. Below we provide a brief review of work from this period. During the second period, from 1984 through the present, researchers expanded on this early work along a variety of dimensions. We review this work in more detail, drawing from the concepts specified in the previous section.

3.1 Early Computational Research on Discovery

Early work on computational approaches to discovery focused on finding empirical regularities such as taxonomies and laws. This was a natural stages of a scientific discipline. Thus, it should require less domain knowledge and permit the use of general heuristics.

Lenat's (1979) AM was one of the earliest discovery systems, operating in the domain of elementary number theory. This domain is unusual when viewed in the light of more recent work, in that one can generate data internally rather than observing them in a real or simulated environment. The user provided AM with an initial taxonomy of mathematical concepts, which it proceeded to extend and revise by mutation. Upon defining a new concept, the system used the definition to generate examples, which it then used to direct the search for other concepts. AM could also posit that two concepts were equivalent even though they had different definitions, as well as notice relations among different concepts. Thus, the system could discover certain classes of qualitative laws, revise its taxonomy, create new terms, and observe examples of these terms. However, it lacked components for experimentation, explanation, prediction, theory formation, and evaluation.

Another early discovery system was Langley, Zytkow, Bradshaw, and Simon's (1983) Bacon, which focused on the induction of numeric laws from experimental data.[3] This program was provided with a set of independent and dependent variables, which it used to carry out simple experiments drawing on simulated data, and which it used to organize results into a taxonomic hierarchy. Once Bacon had gathered data for a given node in its hierarchy, it searched for constant values of dependent terms or relations between independent and dependent terms. In the former case, it augmented the node's description with that constancy; in the latter case, it defined new terms as products or ratios of existing terms and continued the search. The system propagated constant values to higher levels in its hierarchy, where it treated them as dependent values in its search for higher-level numeric laws. Bacon's main contribution was in the area of quantitative discovery and term definition, though it also included user-specified methods for experimentation, taxonomy formation, and observation. Like AM, it contained no explicit components for explanation, prediction, theory formation, or evaluation.

[3] Langley et al.'s approach was influenced by earlier work on function discovery by Huesmann and Cheng (1973) and by Gerwin (1975). Langley, Simon, Bradshaw,

Langley et al. (1983) described two additional systems that address different aspects of the discovery process. Glauber carried out a form of taxonomy formation that also produced simple qualitative laws relating the categories it defined.[4] Stahl formulated simple structural theories of chemical substances based on observed reactions, carrying out a revision process upon encountering anomalous observations that could not be explained by existing theories. Neither system contained explicit methods for experimentation, prediction, or evaluation.

At the level of our framework, AM and Bacon cover similar aspects of the scientific process. Although both systems tackled important aspects of scientific discovery, they also ignored many components of the overall process and thus constituted initial forays rather than integrated models. During the past five years, research on computational approaches to scientific discovery has produced a number of advances over this early work. One can divide these developments into progress in knowledge representation, progress on methods for discovery, and progress on the integration of these methods. In the remainder of this section, we discuss each of these in turn, providing examples from the recent literature.

3.2 Progress on Scientific Knowledge Structures

The most basic advances in machine discovery have involved the representation of observations, laws, models, and theories. Early work assumed simple descriptions of objects and events in terms of numeric attributes or, at best, relations among objects. Qualitative and quantitative representations were entirely separate, and there existed no explicit representation for temporal information. However, a number of recent discovery systems have drawn heavily on Forbus' (1985) work on qualitative process representations. This approach represents events as a sequence of qualitative states, with each state describing an interval of time during which the signs of derivatives remain constant. Forbus' framework also lets one represent theories about processes in qualitative terms and provides mechanisms for making qualitative predictions.

At least four researchers have incorporated this qualitative process representation directly into their discovery systems. For instance, Falkenhainer's Phineas (this volume) uses qualitative data to retrieve and match against promising background knowledge, then forms a new process theory by analogy with this knowledge. O'Rorke, Morris, and Schulenberg (this volume) represent data and theories in a similar form but use anomalies to drive the process of theory revision. Rajamoney's Coast (this volume) uses a qualitative representation for models but uses qualitative anomalies to constrain the experimentation process.[5] Finally, Nordhausen and Langley's IDS (this volume) uses Forbus' formalism to represent both observations and qualitative laws, including ones that involve relations among successive states.

Another representational advance involves the storage of justifications on theories that aid in the processes of theory evaluation and revision. For instance, Thagard and Nowak (this volume) explicitly represent the arguments for and against competing theories, using this information in their evaluation mechanism. In a similar manner, Pazzani and Flower (this volume) make an analogy between theory evaluation and argumentation, proposing the use of explicit arguments and counterarguments in evaluating theories. Rose and Langley (1986) take a related approach in their Stahlp system, indexing observations by the theories they support and retrieving them when anomalies call the theory into question. Rajamoney's Coast (this volume) employs a similar strategy but stores only some of the evidence for a given theory to use during later revisions.

A final representational innovation concerns the role of imagery. Miller (1986) and Tweney (this volume) argue for the central role of imagery in scientific thinking and call for research on computational approaches to this topic. A number of researchers in qualitative reasoning are explicitly working on the problem of spatial reasoning (e.g., Nielsen, 1988), and Shrager's work (this volume) constitutes a novel approach, introducing a representation of scientific knowledge that is grounded in sensory-motor operations. The use of qualitative process formalisms also bears on this topic, in that one can "run" qualitative simulations to "envision" what may follow from given starting conditions. These are only beginnings, but they considerably extend the simplistic mathematics-based schemes that predominated in the early work on discovery.

[4] More recently, Jones (1986) has described an incremental version of GLAUBER that contains explicit components for experimentation, prediction, and

[5] Kulkarni and Simon (this volume) and Karp (this volume) also employ qualitative representations in the design of experiments, but they do not explicitly work in

3.3 Progress on Discovery-Related Activities

In terms of scientific activities, the most impressive advances have occurred with respect to the formation and revision of theories. Falkenhainer's work on analogy describes one approach to theory formation, in which knowledge of other domains is transferred to the one under study. Kulkarni and Simon, O'Rorke et al., and Rajamoney all focus on theory revision, showing how anomalies can lead to modification of an initial theory and its gradual improvement over time. Karp's Hypgene uses similar methods to deal with the related problem of model revision, and Darden (this volume) discusses similar issues in her historical analysis. Rose (1989) describes a unified approach to incrementally revising both theories and observations. Contrasting approaches to theory revision have been proposed that rely on conceptual combination (Holland, Holyoak, Nisbett, & Thagard, 1986; Shrager, 1987), and Shrager's work (this volume) follows this approach in novel directions.

Another area of progress has involved experimentation. Klahr, Dunbar, and Fay (this volume), following upon the theory formation studies of Shrager and Klahr (1986), have carried out detailed studies of the experimentation strategies of humans in understanding complex devices, extending previous work (e.g., Mynatt, Doherty, & Tweney, 1978) in important ways. The computational models of Kulkarni and Simon, Rajamoney, and Karp have all focused on experimentation, and their approaches share some important similarities. Each of their systems makes predictions, notes anomalies, uses the latter to generate alternative hypotheses, and then designs experiments to discriminate among the competitors.

Although a smaller fraction of researchers have focused on empirical discovery than in earlier days, advances have also occurred along this front. One development is the work on "conceptual clustering" by Stepp (1984), Lebowitz (1987), and Fisher (1987), which organizes observations into taxonomies of concepts described at varying levels of abstraction. Another area concerns improved methods for discovering numeric laws, such as those described by Falkenhainer and Michalski (1986), Kokar (1986), and Zytkow (1987). More recently, Nordhausen and Langley (this volume) have reported novel methods in both areas, along with techniques for discovering qualitative laws. Zytkow (this volume) outlines a method for quantitative discovery that takes advantage of domain models to parse numeric laws into useful components. Both

approaches rely on more powerful representations of observations and laws than were used in earlier work. Another line of research by Epstein (1987), Shen (1990), and Sims and Bresina (1989) has continued in the AM tradition, refining Lenat's approach and applying it to new mathematical domains.

Finally, research has also progressed in the area of evaluation. Thagard and Nowak (this volume) describe a method for evaluating the relative quality of two theories in terms of each theory's ability to explain a variety of phenomena. Taking a different approach, Cheeseman (this volume) proposes Bayesian probabilistic criteria for evaluating taxonomies and laws. Both approaches seem likely to find their way into future discovery systems, where they could be used to direct the search for improved laws and theories.

3.4 Progress on Integrated Approaches to Discovery

Another important trend has been the evolution toward *integrated* discovery systems. A number of researchers have combined nontrivial components of the discovery process, producing synergistic effects from their interactions. One can view these efforts as steps along the path toward a complete theory of scientific discovery that describes not only basic activities but also the relations among them.

One relatively complete integration of activities is embodied in Shrager's (1987) IE system, which carried out experiments on simulations of a complex device and formed "mental models" of the device by conceptual combination. The system performed explorations and experiments (both involving prediction) on the simulated device and carried out exercises in order to test the completeness of its theory. Although Shrager was concerned mainly with IE's "view application" method for theory reformulation, the model also included simple versions of analogical theory extension and postdiction.

Another example is Nordhausen and Langley's work, which integrates taxonomy formation, qualitative law discovery, and numeric law discovery. Their IDS system incrementally organizes observed qualitative states into a taxonomic hierarchy and then formulates qualitative laws in terms of temporal relations between classes of states. It also uses these qualitative laws to provide context for numeric relations and to constrain the search for the latter.

A third case is Kulkarni and Simon's Kekada, which integrates theory revision, prediction, experimentation, and evaluation. Their system begins with a partial theory and an anomaly, which Kekada attempts to explain by elaborating the theory. This leads to a number of alternative hypotheses, which the system evaluates by designing and running experiments. If Kekada encounters some new anomaly along the way, it shifts attention and follows this path instead.[6]

These three systems are not the only ones that attempt to integrate aspects of the discovery process, but they provide prototypical examples of this trend. If one compares the above descriptions of IE, IDS, and Kekada with the earlier characterizations of AM and Bacon, the recent progress toward integrated models of scientific discovery becomes apparent. However, it is also clear that much work remains before we arrive at a model that fully integrates even the incomplete set of processes included in the framework from Section 2.

4. Open Issues in Scientific Discovery

In closing, we consider two important aspects of intellectual activity—*embedding* and *embodiment*—that have significant bearing on science but that have not been addressed by existing computational models. Briefly, science takes place in a world that is occupied by the scientist, by the physical system under study, and by other agents, and this world has indefinite richness of physical structure and constraint. Thus the scientist is an embodied agent embedded in a physical and social world.

Embodiment brings to the fore components of scientific behavior that are easily ignored when the model exists entirely within a computer, where all aspects of the environment are controllable, where observation can take place by direct reference to data structures, and where the environment has finite and known complexity. Embedding highlights issues that have been traditionally ignored by models that focus on the intellectual activity of individual scientists rather than on communities. In this section, we consider some results of embedding and embodiment that have generally been ignored in computational models of scientific behavior (see also the critique of Tweney, this volume). We end by discussing some promising approaches toward creating computational accounts for these components.

4.1 External Representations and Research Programs

As scientific domains become increasingly data intensive, external representations come to play a central role in the research process. For instance, notebooks and graphics are widely used in many disciplines as memory aids and, more importantly, as aids to discovery through data organization. In addition, such records help in research planning, in which scientists sequence their activities within the larger scientific context. Several researchers (Darden, this volume; Gorman & Carlson, in press; Tweney, this volume) have studied the use of laboratory notes and records, along with their influence on scientific reasoning. Their analyses suggest that these external records have a major influence on the discovery process. Kulkarni and Simon (this volume) are concerned with programs of research, but they do not model the role of external records in the planning process.

Given the importance of notebooks, graphics, and similar records, it may seem astonishing that none of the existing computational models of discovery incorporate such devices. Part of the reason comes from unrealistic assumptions about the memory and speed of computational systems. For instance, Langley et al.'s Bacon has no need to plot its data since it can retain as many observations as necessary in working memory and scan the data rapidly. Although no computational models of discovery have taken seriously the function of external representations, some work has been done in other areas of cognitive science (e.g., Larkin & Simon, 1987; Shrager, 1989). Also, Shrager's theory of grounded representation (this volume) partially addresses this issue, in that it is designed to operate with external stimuli as well as with internal sensory content.

4.2 Perception and Measurement Instruments

The measurement process alone occupies a major fraction of scientists' time and energy, leaving precious little remaining time for the intellectual activities that we considered in Section 2. However, existing models of scientific discovery are disembodied; they assume immediate

[6] Rajamoney's COAST and Karp's HYPGENE also combine prediction, experimentation, and revision, but they focus on individual steps in this process rather than

itly separate the external setting from the agent's internal knowledge, the environment is sufficiently constrained that issues of attention and perception are avoided. In addition to measuring simple quantities, scientists must also connect observables to theoretical terms if the latter are to be operational. Recent research on attention in concept learning (Billman & Heit, 1988; Gennari, 1989) has started to address some of these issues, but much more remains to be done.

Moreover, even the earliest histories of discovery involve some forms of instrumentation. Many authors (e.g., Feyerabend, 1975; Giere, 1988; Hacking, 1983) have noted the importance of measurement instruments in the scientific process, but computational models have ignored this aspect of research. As with perception, this oversight is understandable, in part because instrument construction is largely a physical phenomenon that is difficult to model without solving difficult problems in robotics or building rich simulations. One approach that shows some promise is Nordhausen and Langley's (this volume) method for postulating *intrinsic properties*, which provides a method for computing features of new objects based on their behavior in familiar qualitative histories. Effectively, these abstract histories describe "instruments" that let one measure properties like boiling point or specific heat.

4.3 Laboratories, Collaboration, and Communication

Most modern science is too large and too expensive an undertaking for an independent researcher to succeed, making it essential that scientists collaborate. Although there are many alternative organizations for joint research, the most common is the laboratory, in which a small number of researchers collaborate on a small set of problems. Laboratories generally exist at a single location and include scientists at different levels of expertise, from students to senior researchers. In addition, different laboratories often work on the same or closely related problems. In some cases, this work is competitive, but in other cases there is significant cooperation, with division of labor and open interactions.

Collaboration of any sort requires some form of communication among scientists, and it takes no statistical sophistication to conclude that scientists spend much of their time talking, reading, and writing. These sorts of communication provide another example of external representations in which graphics, mathematical expression, and language play before anything reaches the formal scientific literature. Tweney (this volume) has argued that Faraday enriched his understanding of one domain from his ongoing research in other domains, but such enrichment is surely is not restricted to the mental activities of individual scientists. Formal communication is essential for the broader dissemination of ideas, making reading and writing central scientific activities.

Existing computational models of discovery have avoided the collaborative and communicative aspects of scientific research, focusing on individual scientists' behavior and ignoring group interactions. This was a natural development, given the traditional focus of cognitive science on the cognitive processes of individuals. However, the social organization of science in the laboratory and in broader contexts has a major influence on the nature of science, and future modeling efforts should move toward incorporating aspects of this structure.[7]

4.4 Toward a Fuller Computational Account of Discovery

In summary, actual science occurs in the context of a physical world and in the context of other agents, but existing computational accounts of discovery have avoided these major issues. The reason for this bias is straightforward. The methods and theories of cognitive science were originally designed to model individual cognition, and the computational work on discovery has relied heavily on these tools. A deeper understanding of embodiment will require considerable research in AI and cognitive psychology, and the embedded nature of science awaits additional work in sociology, anthropology, and psychology. Nevertheless, some preliminary results hold out hope for advances in these areas.

For instance, the active research area of "distributed artificial intelligence" focuses on understanding the ways that multiple agents can interact in communities. Several collections are available on this topic (see Gasser & Huhns, 1989; Huberman & Hogg, 1988), and researchers

[7] Excellent collections have recently appeared on the sociology of scientific practice and knowledge (see Fuller, De May, Shinn, & Woolgar, 1989). Readers of this chapter will be particularly interested in a special issue of *Social Studies of Science* (volume 19, number 4), in which several authors respond to Slezak (1989), who argues that the success of BACON and similar discovery programs "[provide] dramatic confirmation [of the view that] there are principles of rationality and a

in this field explicitly draw upon results in the social sciences, especially from economics and scientific reasoning. There is also hope that psycho-anthropological approaches (e.g., Latour & Woolgar, 1979; Lynch, 1985; Pickering, 1984) will explain certain social aspects of science, especially the role of communication. However, to date these accounts have been descriptive rather than computational. The literature on distributed artificial intelligence also deals with issues of communication but focuses on the nature of the information passed rather than on the processes of individual agents acting in the community. Overall, there has been little computational work on the communicative interactions of agents with one another. Thagard and Nowak's work (this volume) on the acceptance of revolutions most closely speaks to the issues of interactions among researchers, but their paradigm does not model the richness and detailed functions of scientific communication.

Research on embodied agents has also made progress, not only in traditional approaches to robotics but also in the interface between AI, machine learning, and robotics. For instance, Laird, Yager, Tuck, and Hucka (1989) describe a system that improves its ability to use a robot arm with experience. The work of Iba and Langley (1987) on motor learning provides an additional example of this encouraging trend. More relevant to scientific discovery are recent attempts (Zytkow, Zhu, & Hussam, in press) to employ AI methods to control robotic equipment for chemical experimentation. In addition, some researchers (see Shrager, this volume) have taken perception as a central problem and have attempted to explain complex intellectual activity in terms of sensation and action. Other researchers have even attempted to deal with the issues of physical and social environments simultaneously, as Cohen, Greenberg, Hart, and Howe (1989) have done in their work on cooperative fire fighting in a simulated (burning) forest.

We believe that an important source for models of embedding and embodiment in science will come from an unexpected direction: the developmental psychology of *socialization*, which studies the ways in which a child learns to become a part of his or her culture (e.g., Bruner, 1985; Kuhn, Amsel, & O'Loughlin, 1988; Vygotsky, 1962). Insights into this process may provide hypotheses about the paths through which graduate students and junior scientists become members of their scientific

community—mastering the ways of thinking, operating, and communicating that constitute the institution of science.[8]

5. Conclusion

In effect, this chapter has attempted to define a new field of study—the computational modeling of scientific behavior. Despite its relatively recent development, this research area has already made significant progress on issues that philosophers of science have traditionally ignored. In particular, the field has emphasized the nature of *discovery* rather than evaluation, and it has dealt with the *processes* that underlie science as well as the representation of knowledge. The result has been a rapidly growing set of computational models that deal with many facets of the scientific enterprise.

Although the existing models are best viewed as embodying tentative hypotheses about the nature of science, it is also clear that the past decade has seen real progress. Current systems still ignore many important aspects of discovery and theory formation, but idealizations are a central part of science; we should no more expect our computer simulations to account for *every* aspect of discovery than we expect our physical or chemical theories to explain every aspect of the physical world. What we can expect is incremental progress toward fuller models and deeper understanding, and that is precisely what has occurred in the developing computational "science of science."

The past few years have seen notable developments, not only in the representations and processes used to model scientific discovery and theory formation but also in their integration into a coherent framework. We will not make specific predictions about the outlook for extending the computational paradigm into the more difficult areas of embedded and embodied science. However, the paths toward these goals seem lined with fertile research questions waiting to be addressed. Progress along these paths will certainly tax our existing theories and methodology, but it should also bear rich rewards.

[8] Luhrmann's (1989) insightful psychoethnography of British witchcraft provides a carefully researched example of a sort of socialization that she calls "interpretive drift." The analogy between becoming a scientist and becoming a witch runs more deeply than one might think. Both deal with belief and action, and both have significant rites of passage. Most of the structures and activities that we have identified as typical of science apply equally well to witchcraft, and even to

Acknowledgements

We owe thanks to many colleagues for discussions that led to the ideas in this chapter. In particular, Mike Morgan suggested that we develop an encompassing vocabulary; Rik Belew and Michael Gorman encouraged us to think more about the social aspects of science; and Deepak Kulkarni provided helpful comments on an earlier draft.

References

Billman, D., & Heit, E. (1988). Observational learning from internal feedback: A simulation of an adaptive learning method. *Cognitive Science*, *12*, 587–626.

Bruner, J. (1985). *Child's talk.* New York: W. W. Norton.

Carey, S., & Wiser, M. (1983). When heat and temprature were one. In D. Gentner & A. L. Stevens (Eds.), *Mental models*. Hillsdale, NJ: Lawrence Erlbaum.

Cohen, P. R., Greenberg, M. L., Hart, D. M., & Howe, A. E. (1989). Trial by fire: Understanding the design requirements for agents in complex environments. *AI Magazine*, *10*, 32–48.

Epstein, S. L. (1987). On the discovery of mathematical theorems. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 194–197). Milan, Italy: Morgan Kaufmann.

Falkenhainer, B. C., & Michalski, R. S. (1986). Integrating quantitative and qualitative discovery: The Abacus system. *Machine Learning*, *1*, 167–402.

Falkenhainer, B. C., & Rajamoney, S. (1988). The interdependencies of theory formation, revision, and experimentation. *Proceedings of the Fifth International Conference on Machine Learning* (pp. 353–366). Ann Arbor, MI: Morgan Kaufmann.

Feyerabend, P. (1975). *Against method.* London: Verso.

Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, *2*, 139–172.

Forbus, K. D. (1985). Qualitative process theory. In D. G. Bobrow (Ed.), *Qualitative reasoning about physical systems.* Cambridge, MA: MIT Press.

Fuller, S., De May, M., Shinn, T., & Woolgar, S. (1989). *The cognitive turn: Sociological and psychological perspectives on science.* Boston: Kluwer Academic Publishers.

Gasser, L., & Huhns, M. N. (Eds.). (1989). *Distributed artificial intelligence* (Vol. 2). San Mateo, CA: Morgan Kaufmann.

Gennari, J. H. (1989). Focused concept formation. *Proceedings of the Sixth International Workshop on Machine Learning* (pp. 379–382). Ithaca, NY: Morgan Kaufmann.

Gerwin, D. (1975). Information processing, data inferences, and scientific generalization. *Behavioral Science, 19*, 314–325.

Gholson, B., Shadish, W. R., Neimeyer, R. A., & Houts, A. C. (1989). *Psychology of science: Contributions to metascience*. Cambridge: Cambridge University Press.

Giere, R. N. (1988). *Explaining science.* Chicago: University of Chicago Press.

Gorman, M. E., & Carlson, B. W. (in press). Interpreting invention as a cognitive process: Alexander Graham Bell, Thomas Edison, and the telephone, 1875–1878. *Science, Technology, and Human Values*.

Hacking, I. (1983). *Representing and intervening.* Cambridge: Cambridge University Press.

Holland, J. H., Holyoak, K. J. , Nisbett, R. E. , & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Huberman, B. A., & Hogg, T. (1988). The behavior of computational ecologies. In B. A. Huberman (Ed.), *The ecology of computation* (pp. 77–115). New York: North-Holland.

Huesmann, L. R. , & Cheng, C. M. (1973). A theory for the induction of mathematical functions. *Psychological Review, 80*, 126–138.

Iba, W., & Langley, P. (1987). A computational theory of motor learning. *Computational Intelligence*, *3*, 338–350.

Jones, R. (1986). Generating predictions to aid in the scientific discovery process. *Proceedings of the Fifth National Conference on Artificial Intelligence* (pp. 513–517). Philadelphia, PA: Morgan Kaufmann.

Kokar, M. M. (1986). Determining arguments of invariant functional descriptions. *Machine Learning*, *1*, 403–422.

Kuhn, D., Amsel, E., & O'Loughlin, M. (1988). *The development of scientific thinking skills.* New York: Academic Press.

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago, IL: University of Chicago Press.

Laird, J., Yager, E. S., Tuck, C. M., & Hucka, M. (1989). Learning in teleautonomous systems using Soar. *Proceedings of the 1989 NASA Conference on Space Telerobotics*. Pasadena, CA.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge.* Cambridge: Cambridge University Press.

Lakatos, I. (1976). *Proofs and refutations: The logic of mathematical discovery.* J. Worrall & E. Zahar (Eds.). Cambridge: Cambridge University Press.

Langley, P., Simon, H. A., Bradshaw, G. L., & Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative processes* (pp. 251–283). Cambridge, MA: MIT Press.

Langley, P., Zytkow, J., Bradshaw, G., & Simon, H. A. (1983). Three facets of scientific discovery. *Proceedings of the Eighth International Joint Conference on Artificial Intelligence* (pp. 465–468). Karlsruhe, West Germany: Morgan Kaufmann.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*, 65–99.

Latour, B., & Woolgar, S. (1979). *Laboratory life: The social construction of scientific facts.* London: Sage.

Lebowitz, M. (1987). Experiments with incremental concept formation: Unimem. *Machine Learning*, *2*, 103–138.

Lenat, D. (1979). On automated scientific theory formation: A case study using the AM program. In J. Hayes, D. Michie, & L. I. Mikulich (Eds.), *Machine intelligence* (Vol. 9). New York: Halstead Press.

Luhrmann, T. M. (1989). *Persuasions of the witch's craft.* Cambridge, MA: Harvard University Press.

Lynch, M. (1985). *Art and artifact in laboratory science: A study of shop work and shop talk in a research laboratory.* London: Routledge & Kegan Paul.

Miller, A. I. (1986). *Imagery in scientific thought.* Cambridge, MA: MIT Press.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1978). Consequences of confirmation and disconfirmation in a simulated research environment. *Quarterly Journal of Experimental Psychology*, *30*, 395–406.

Nielsen, P. (1988). A qualitative approach to mechanical constraint. *Proceedings of the Seventh National Conference on Artificial Intelligence* (pp. 270–274). Saint Paul, MN: Morgan Kaufmann.

Pickering, A. (1984). *Constructing quarks.* Chicago: University of Chicago Press.

Popper, K. (1965). *Conjectures and refutations: The growth of scientific knowledge* (2nd ed.). New York: Basic Books.

Rose, D. (1989). Using domain knowledge to aid in scientific theory formation. *Proceedings of the Sixth International Conference on Machine Learning* (pp. 272–277). Ithaca, NY: Morgan Kaufmann.

Rose, D., & Langley, P. (1986). Chemical discovery as belief revision. *Machine Learning*, *1*, 423–451.

Shen, W. M. (1990). Functional transformation in AI discovery systems. *Artificial Intelligence*, *41*, 257–272.

Shrager, J. (1987). Theory change via view application in instructionless learning. *Machine Learning*, *2*, 247–276.

Shrager, J. (1989). Reinterpretation and the perceptual microstructure of conceptual knowledge: Cognition considered as a perceptual skill. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society* (pp. 876–883). Ann Arbor, MI: Lawrence Erlbaum.

Shrager, J., & Klahr, D. K. (1986). Instructionless learning about a complex device: The paradigm and observations. *International Journal of Man-Machine Studies*, *25*, 153–189.

Sims, M. H., & Bresina, J. L. (1989). Discovering mathematical operator definitions. *Proceedings of the Sixth International Conference on Machine Learning* (pp. 308–313). Ithaca, NY: Morgan Kaufmann.

Slezak, P. (1989). Scientific discovery by computer as empirical refutation of the strong programme. *Social Studies of Science*, *19*, 563–600.

Stepp, R. (1984). *Conjunctive conceptual clustering: A methodology and experimentation.* Doctoral dissertation, Department of Computer Science, University of Illinois, Urbana-Champaign.

Suppe, F. (1977). *The structure of scientific theories* (2nd ed.). Urbana, IL: University of Illinois Press.

Vygotsky, L. S. (1962). *Thought and language.* (E. Hanfmann & G. Vakar, trans.). Cambridge, MA: MIT Press.

Weld, D. S. , & de Kleer, J. (Eds.). (1990). *Qualitative reasoning about physical systems.* San Mateo, CA: Morgan Kaufmann.

Zytkow, J. M. (1987). Combining many searches in the Fahrenheit discovery system. *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 281–287). Irvine, CA: Morgan Kaufmann.

Zytkow, J. M., Zhu, J., & Hussam, A. (in press). Automated discovery in a chemistry laboratory. *Proceedings of the Eighth National Conference on Artificial Intelligence.* Cambridge, MA: AAAI Press.

Contents