HERBERT A. SIMON, PATRICK W. LANGLEY,
AND GARY L. BRADSHAW

SCIENTIFIC DISCOVERY AS PROBLEM SOLVING*

The question to be addressed in this paper is whether we need a special theory to explain the mechanisms of scientific discovery, or whether those mechanisms can be subsumed as special cases of the general mechanisms of human problem solving. One of the authors has previously published several papers arguing for the latter position.[1] The main evidence adduced in those papers for the thesis that scientific discovery is problem solving was the behavior of some computer programs that, using simple problem-solving heuristics and selective search, were capable of discovering patterns in simple sequences of symbols.[2] Much stronger evidence has now been provided by the performance of D. B. Lenat's AM program,[3] which discovers mathematical concepts and conjectures theorems, and P. W. Langley's BACON programs,[4] which discover invariants in bodies of empirical data. It is a main purpose of this paper to review this new evidence and its implications for the theory of scientific discovery.

Of course there are several respects in which scientific discovery is obviously different from other instances of problem solving. First, scientific inquiry is a social process, often involving many scientists and often extending over long periods of time. Much human problem solving, especially that which has been studied in the psychological laboratory, involves a single individual working for a few hours at most.

A second way in which scientific inquiry differs from much, but not all, other problem solving is in the indefiniteness of its goals. In solving the Missionaries and Cannibals puzzle, we know exactly what we want to achieve: we want a plan for transporting the missionaries and cannibals across the river in the available small boat without any casualties from drowning or dining. Some scientific discovery is like that: The mathematicians who found a proof for the Four-color Theorem knew exactly what they were seeking. So did Adams and

Synthese **47** (1981) 1–27. 0039–7857/81/0471–0001 $02.70

Leverrier when they detected Neptune while searching for a celestial object that would explain anomalies in the orbits of the other planets.

In most scientific inquiry, however – and especially in what Kuhn has called "revolutionary science" – the targets are less sharp. What was Darwin seeking that was more definite than a way of putting order into the complex, confusing data of biological diversity and fossil geology? Toward the end of the eighteenth century, just prior to the work of Lavoisier, how would one have defined the "problem of combustion"?

In spite of these differences between scientific inquiry and other problem solving, it is quite possible that the component processes, which when assembled make the mosaic of scientific discovery, are not qualitatively distinct from the processes that have been observed in simpler problem-solving situations. Solving complex problems generally involves decomposing them into sets of simpler problems and attacking these. It could well be the case (and we will argue that it in fact is) that the component problem solving involved in scientific discovery has no special properties to distinguish it from other problem solving.

Our paper will be concerned with describing and explaining scientific discovery rather than with providing a normative theory of the process. Indeed, the very possibility of a normative theory has been challenged by many philosophers of science.[5] However, if we succeed in producing a credible explanation of discovery, that explanation will itself also constitute a first approximation to a normative theory. Especially is this true if the explanation is constructive – if it exhibits a set of processes that, when executed, actually make scientific discoveries. The explanation we shall propose is of this constructive kind.

The paper is divided into three major parts. The first part will comment on the anatomy of scientific discovery – perhaps it would be more modest to call it the taxonomy. There is no unitary activity called "scientific discovery"; there are activities of designing experiments, gathering data, inventing and developing observational instruments, formulating and modifying theories, deducing consequences from theories, making predictions from theories, testing theories, inducing regularities and invariants from data, discovering theoretical constructs, and others. All of these activities that make up the scientific enterprise must be put in perspective, and that is what we shall try to do in the first part of the paper.

In the second part of our paper we will describe a computer program, called BACON.4, that constitutes a constructive theory of some important kinds of scientific discovery. We will support this claim by examples of the performance of BACON.4 when it is placed in various historical situations, that is, when it is confronted with the same sorts of empirical data that were available to physicists and chemists during certain critical episodes in the eighteenth and nineteenth century histories of those sciences.

In the third part of the paper, we will describe, more briefly, a program, AM, that makes discoveries of a rather different kind from BACON.4's, using theory-driven instead of data-driven processes. We will also comment more generally on what part of the spectrum of scientific activity is illuminated by the BACON.4 and AM programs.

PROCESSES OF DISCOVERY

The scientific enterprise is dedicated to the extension of knowledge about the external world. It is usually conceived as being made up of three main kinds of interrelated activities: gathering data, finding parsimonious descriptions of the data, and formulating and testing explanatory theories. Sometimes the second (description) and third (explanation) categories are merged. Usually these activities are conceived as occurring in cyclical fashion. Theories are formulated, from these predictions are made, data are gathered, and the theories are tested by confronting their predictions with the data. Failure of data to support theories leads, in turn, to the formulation of new theories. It is generally agreed, however, that the actual sequences of events are less regular. Data may be gathered without clear theoretical preconceptions, and theories may be retained, especially in the absence of viable alternatives, even after some of their predictions have been disconfirmed.

Some Comments on Taxonomy

Whether or not data gathering, description, explanation, and theory testing are strictly cyclical, each of these activities can be subdivided further. Data may be gathered by observation of natural events or by 'producing' phenomena through experimentation. If the data are to be obtained by experimenting, the experiments must be designed; and in the cases both of experiment and observation, data-gathering instruments must be invented and improved.

Data are seldom reported in wholly raw form. They must usually be digested and summarized. Summarizing requires detecting regularities and invariants – that is, developing *descriptive generalizations*, which may or may not be dignified by the label 'theory'.

Descriptive generalizations or theories may, in turn, be derivable from *explanatory theories*. We will have more to say presently about how a line might be drawn between these two categories of theory. Whether or not there is, indeed, such a boundary, science includes a whole collection of 'theoretical' activities: inventing theories, deriving theorems from them (including predictions and derivations of descriptive from explanatory theories) as well as directly predicting observations.

Finally, another collection of activities comes under the general heading of *testing theories*. Here are included the whole range of statistical techniques for comparing theoretical statements with data and making judgments of 'significance'.

The *diffusion* of scientific discoveries and expository writing about them are usually regarded as meta-activities of science. Yet we must remember that Mendeleev discovered the periodic table while planning the arrangement of topics for an elementary chemistry textbook.

Strong and Weak Methods in Discovery

By and large it is characteristic of all of these activities of scientific inquiry – with some partial exception for hypothesis testing – that they are usually carried out in a relatively unsystematic and only partly organized way. There is no powerful factory method – no assembly line – for the manufacture of scientific truth. In fact, indications in any domain of science that problem solving can be accomplished systematically and routinely causes that activity to be classified as 'mere' development or application rather than basic scientific inquiry. Even if the activity falls far short of complete systematization and routinization, if its 'do-ability' is nearly predictable, it falls under Kuhn's rubric of 'normal' science rather than the more prestigious 'revolutionary' science. In fact, methodologists of science sometimes hint that the fundamentality of a piece of scientific work is almost inversely proportional to the clarity of vision with which it can be planned.

Similar gradations and distinctions in degree of systematization are made in theories of problem solving. Expert problem solving in any domain is characterized by the systematic application of powerful methods, each adapted to a recognizable subdomain. The expert does not have to seek and search; he recognizes and calculates. Problem solving by novices, on the other hand, is characterized by tentativeness and uncertainty. The novice frequently does not know what to do next; he searches and tests; he uses crude and cumbersome methods because he is not familiar with the powerful tools of the expert.[6]

It is understandable, if ironic, that 'normal' science fits pretty well the description of expert problem solving, while 'revolutionary' science fits the description of problem solving by novices. It is understandable because scientific activity, particularly at the revolutionary end of the continuum, is concerned with the discovery of new truths, not with the application of truths that are already well known. While it may incorporate some expert techniques in the manipulation of instruments and laboratory procedures, it is basically a journey into unmapped terrain. Consequently, it is mainly characterized, as is novice problem solving, by trial-and-error search. The search may be highly selective – the selectivity depending on how much is already known about the domain – but it reaches its goal only after many halts, turnings, and backtrackings.

By necessity, problem solving in novel domains makes use of *weak methods* – problem solving techniques of quite general application whose generality is assured by the fact that they do not use or require much prior knowledge of the structure of the problem domain. Problem solving in well-understood domains uses *strong methods* – powerful techniques that are carefully tailored to the specific structure of the domain to which they are applied. Of course strong methods are almost always more powerful than weak methods, but they are simply not available, by definition, when entirely new territory is to be explored.

The commonest weak problem solving methods are generate-and-test, heuristic search, and means-ends analysis. The generate-and-test method employs one process to produce candidate problem solutions while another process tests each candidate to determine whether it meets the criteria of solution. Clearly, this method is applicable, in principle, to any problem domain. Clearly also, it will rarely produce solutions if the generator operates on a 'try everything' or 'search randomly' basis, without any information about the problem domain.

The heuristic search method generates new solution candidates by modifying candidates already tried. It can be more efficient than the generate-and-test method to the extent that information can be extracted from the search to guide the direction of modification. Evaluation of solution candidates on some scale of better or worse can provide one such search criterion.

Means-ends analysis is a special form of heuristic search that guides the search at each step by comparing a solution candidate with the criteria for solution, detecting differences between them, and modifying the candidate in a way that is aimed at reducing or removing one or more of the differences. Again, the efficacy of such a scheme depends on what knowledge the system has of the association between specific differences and specific operations for removing them.

Weak methods exploit as little or as much knowledge about the structure of the problem space as is available to them. When little knowledge is available, they are more or less equivalent to, and as inefficient as, random search through the problem space. If a great deal of knowledge is available, especially to the means-ends method, they may home in on solutions with little extraneous search.

Two and one-half decades of research in artificial intelligence and cognitive simulation have shown that weak methods can solve problems in domains that are regarded as difficult for intelligent human problem solvers, and can do so without intolerable amounts of search.[7] The programs for scientific discovery on which we shall comment, and specifically the BACON and AM programs, employ heuristic search and other weak methods as their problem solving tools.

Data-Driven and Theory-Driven Science

Since theories encourage the acquisition of new data, and vice versa – as surely as hens engender eggs and eggs hens – scientific discovery can enter the cycle of scientific activity at any point. In the contemporary literature of the philosophy of science, with its (mistaken) emphasis on theory testing as the quintessential scientific activity, the tale usually begins with a theory. The theory, emanating from the brain of Zeus or some other unexamined source, calls for testing, which demands, in turn, that appropriate data be obtained through observation or experiment. In this scheme of things, the discovery process may be described as *theory driven.*

Especially when a theory is expressed in mathematical form, theory driven discovery may make extensive use of strong methods associated with the mathematics or with the subject matter of the theory itself. The discovery procedures will not be general in this case, but will be tailored to the problem domain.

The converse view, which we may call Baconian induction, takes a body of data as its starting point and searches for a set of generalizations, or theory, to describe the data parsimoniously or to explain them. Usually such a theory takes the form of a precise mathematical statement of the relations among the data. In this scheme of things, the discovery process may be described as *data driven.* Data driven discovery will perforce make use of weak methods, but with the compensating advantage that the methods are quite general in their domain of application. Moreover, we must remember that 'weak methods' will still usually be far more powerful than blind trial-and-error search.

Both data-driven and theory-driven processes give partial views of the scientific enterprise. It is easy to find in the history of science examples of discoveries of first importance that fit each view. Published accounts of the history tend most often to cut the cycle so as to emphasize theory as the first mover – for example Lavoisier's refutation of the phlogiston theory with his careful experiments on mercuric oxide, the disconfirmation of the classical theory of the ether by Michelson and Morley, the test of general relativity by observations on the solar eclipse of 1921, and many others.

On the other hand, the historical records of Mendeleev's discovery of the periodic table and Balmer's of the formula for the hydrogen spectrum reveal both of these to be clearcut cases of data-driven Baconian induction.[8] They could not have been otherwise, since the regularities discovered were, at the time of their discovery, purely descriptive generalizations from the data, lacking any theoretical motivation.

Francis Bacon himself said, in the *Novum Organum*,

Those who have treated the sciences were either empirics or rationalists. The empirics, like ants, lay up stores, and use them; the rationalists, like spiders, spin webs out of themselves; but the bee takes a middle course, gathering up her matter from the flowers of the field and garden, and digesting and preparing it by her native powers.

There is no question, then, of choosing between data-driven and theory-driven discovery. Rather, what is needed is an understanding of how each of these processes can occur. In the next main section of this paper, we will address one part of the task: we will describe the computer program, BACON, which as its name implies is a system capable of making scientific discoveries by induction on bodies of data. In the final section of the paper, we will consider more briefly AM, a system that can do theory-driven discovery.

Description and Explanation

The distinction, alluded to earlier, that is commonly made between *descriptions* and *explanations* is a matter of degree rather than a dichotomy. Generalizations are viewed as 'mere' descriptions to the extent that they stick close to the data, stating one or more relations among the observable variables. Thus, Kepler's Laws, expressed entirely in terms of the shapes of the planetary orbits, the areas over which they sweep, their radii, and their times, are regarded as descriptions rather than explanations of the planetary motions. On the other hand, Newton's Law of Universal Gravitation, $M_1a_1 = gM_1M_2/d^2$, from which Kepler's Laws can be derived, is regarded as an explanation of the latter and, by transivity, of the phenomena they describe. In the Law of Gravitation, the acceleration, a_1, and the distance between the objects, d, are observables, but the masses, M_1 and M_2, are theoretical terms whose values must be inferred indirectly from those of the observables.[9]

Newton's Law, then, is regarded as an explanation both because it contains theoretical terms and because Kepler's descriptive laws can be deduced from it. These seem to be two of the criteria – the presence of theoretical terms and the derivability of more specialized, descriptive consequences – that underlie our intuitions of the distinction between explanatory and descriptive theories.

Although the distinction is intuitively plausible, it does not hold up terribly well under formal examination. For example, Kepler's Third Law can be stated in the form, $S = K$, where K is a constant; whereas $S = P^2/d^3$, where P is the period of the orbit and d is the distance of the planet from the Sun. However, S is then a theoretical term, for it is not observed directly but is computed by the equation given above from the values of P and d, which are observed. Consequently, if we take the presence of a theoretical term in a law as the test for its being explanatory, we would have to classify Kepler's Third Law as explanatory, a classification that does not match very well with our intuitions.

Without trying to arrive at an immediate judgment, let us carry the story a step farther. If a pair of blocks, X and Y, is connected by a spring, the spring stretched to various lengths and then released, and the initial accelerations of the blocks measured, we can discover the invariant relation, $a_X/a_Y = -K_{XY}$, where the a's are accelerations, as before, and K is a theoretical term that is constant. With a new pair of blocks, W and Z, a similar invariance will be found, possibly with a new constant, K_{WZ}. Now we might seek to state the invariance in another way by attributing to each block, X, an invariant property, M_X, and asserting the law, $M_X a_X = -M_Y a_Y$, or what is equivalent, $M_Y/M_X = K_{XY}$. Expressing the ratios of the accelerations in terms of the postulated masses yields a stronger law than the first form of the invariance, for it implies transitivity of these ratios over the whole set of blocks – i.e., that $K_{XZ} = K_{XY}K_{YZ}$. The first form of the invariance is expressed in terms of the ratio of the accelerations, while the second is expressed as a relation between these accelerations and the inertial masses of the individual blocks. Yet, there are theoretical terms in both forms of the law: the K's in the first and the M's in the second. As a matter of fact the number of different values of the theoretical terms (constants) in the first form of the law is n^2 for a set of n blocks, but for the second form of the law it is only n.

There is a difference, however, between the ratios of accelerations, K_{XY}, and the masses, M_X. The former can be computed directly as ratios of observables, while the latter cannot. The *existence* of the mass numbers must be postulated, and the mass of one of the blocks taken as the standard; then the values of the remaining masses can be computed from the ratios of the accelerations. The same distinction between two kinds of theoretical terms can be found by comparing the directly computable constant in Kepler's Third Law with the mass numbers that appear in Newton's gravitational law. In both examples, the explanatory law is slightly more parsimonious than the descriptive law; it also makes use of a more sophisticated kind of theoretical term, introducing into the picture a new property that is not obtainable by taking simple functions of the observables. We will call invariants that are introduced in this way *intrinsic properties.* And from the two examples we have examined, we can conjecture that the introduction of intrinsic properties gives a generalization an explanatory rather than merely descriptive flavor.

These remarks fall far short of providing a formal basis for a distinction between explanations and descriptions, but they provide a background for understanding the fact that in the BACON.4 program, to be described in the next section, different mechanisms are used to introduce the two different kinds of theoretical terms that we have identified here. All of these mechanisms remain in the category of weak methods, so that they do not reduce the generality of BACON.4 nor introduce any great element of complexity. Before we leave the subject of explanation, however, we should comment on a third viewpoint on this topic which we have not mentioned.

Explanation and Causation

We have said that explanatory power may be attributed to a law if it leads deductively to correct predictions of empirical data or if it postulates intrinsic properties. Sometimes, on the other hand, laws are said to be explanatory if they give a *causal* account of the phenomena. Since the correct explication of the notion of causality is still a topic of discussion and dispute in philosophy,[10] we will again proceed informally. Fortunately, most of the philosophical difficulties relate to the attribution of causes in non-experimental situations, while most of our data and theories (except for Kepler's law) are derived from experiment, where matters are relatively straightforward.

If the value of a dependent variable changes when the value of an independent variable is altered, we say that the latter change causes the former. When we replace one of the blocks in the conservation or momentum experiment with a heavier one, the corresponding acceleration decreases. Hence, we say that inertial mass is a deterrent (i.e., a negative cause) to acceleration. We do *not* make the converse inference: that decreasing the acceleration will cause the mass to increase. The asymmetry between the variables appears to arise from the fact that we have a way of intervening in the situation directly to change the masses, but no way of intervening directly to change the ratio of the accelerations. In any event, if we associate explanation with causality, we can regard the law of conservation of momentum as a causal law.

Ohm's law provides a similar example. In a circuit with a battery and a resistance wire, we can replace one battery with another, and one resistance wire with another, measuring the current (the dependent variable) each time. In the theory we derive from these data, the voltage of the battery and the resistance of the wire are the causes that determine the amount of current that flows. The asymmetry is genuine, since we have no physical way of causing the battery or the wire to change by varying the current directly. Ohm's law, by this criterion, is not merely descriptive; it is an explanatory law.

What shall we say about Kepler's Third Law and the Law of Universal Gravitation? In the former, there seems to be a perfect symmetry between distance and period of orbit; neither seems to take precedence in causing the other. Of course in our age of Sputniks, we can turn the situation into an experimental one. We launch a satellite with sufficient energy to put it into orbit at a specified height above the earth. At that altitude, it will orbit with a period given by Kepler's Law. Does the law now provide a causal explanation of the period?

In the case of Newton's Law, we instinctively feel an asymmetry. We read $M_1a_1 = gM_1M_2/d^2$ from right to left and not from left to right. We think of the masses and their distances as causing the acceleration (via the force field *caused* by the gravitational mass), and not vice versa. This reaction is perhaps not unrelated to the fact that the acceleration contains an implicit reference to future velocity and position; hence, if we read the equation from right to left, the causal arrow will agree with the arrow of forward movement through time, while if we read it in the other direction, it will reverse that arrow.

We will not try to resolve here the philosophical issues raised by these examples. Our purpose in mentioning them is to point out that if we take the assertion of causality to be the criterion for a law's being explanatory, then all laws derived from experimental manipulations as well as laws that involve temporal asymmetry can be given explanatory interpretations.

Nor do we claim that we have explicated, in this section and the preceding one, all of the criteria that might be advanced for distinguishing explanatory from descriptive laws. In particular, we have said nothing about laws that *reduce* a description at one level of aggregation to a description at a lower level – the reduction of the laws of thermodynamics, for example, to the laws of statistical mechanics, or the laws of chemistry to quantum mechanics and the laws of atomic physics. There would be general agreement that all such laws are explanatory, but not that *only* such laws are explanatory.

BACON: A DATA-DRIVEN DISCOVERY SYSTEM

We are now ready to return to our main topic: a demonstration that scientific discovery – at least some kinds of scientific discovery – can be accomplished by an inductive process that is driven by empirical data. The demonstration takes the form of reporting the results of experiments with a computer program, BACON.4, that discovers empirical laws using data-driven weak methods. BACON.4 is able to discover both descriptive and explanatory laws.

BACON.4 employs a small set of data-driven heuristics to detect regularities in numeric and nominal data. These heuristics, by noting constancies and trends, cause BACON.4 to formulate hypotheses, define theoretical terms, postulate intrinsic properties, and propose integral relations (common divisors) among quantities. The BACON.4 heuristics do not depend on the specific properties of the particular problem domains to which they have been applied but appear to be general techniques applicable to discovery in diverse domains.

BACON.4's Heuristics

BACON.4 is programmed as a production system whose components are computer instructions called *productions*. Each production consists of two parts: a *condition* part and an *action* part. The conditions are tests that are performed on BACON.4's data base. Some of the actions are arithmetic operations on the data base to compute functions of the data previously developed; other actions bring in new data (i.e., 'perform experiments'). Whenever the conditions of any production are satisfied by the data in the data base, the actions of the production may be carried out. A set of simple rules adjudicates conflicts among productions when two or more seek to execute at the same time; the content of these conflict resolving rules will not concern us. BACON.4's heuristics fall into four classes, which we discuss in detail below.

Detecting Covariance. The nature of the data BACON.4 uses and the way it operates is most easily explained by working through some simple examples. In Table 1 is shown a set of data that might have been gathered by Kepler. Each row of the table except the last one represents an individual observation, and the entire table represents a set of observations made on the planets. Each observation consists of the value of a nominal variable, which identifies the planet, and of two numeric variables, its distance from the Sun, D, and the period of its orbit, P. All of these variables are observables. (The names of the planets could be replaced by descriptions that identified them uniquely by their color, brightness, location, and so on.)

To extract Kepler's Third Law from these data, BACON.4 examines the column for each observable (D and P) to determine whether the values in the column are all identical. Finding that none are in this case, it arranges the observations monotonically according to the values of one variable (say, D) and then examines the columns of the other variables for monotonicity. In this case, since period of orbit, P, increases monotonically with D, BACON.4 forms the ratio, D/P, but finds that it is not invariant. Since this ratio varies inversely with D, BACON.4 now computes the product, $(D/P) \times D = D^2/P$.

TABLE I
Induction of Kepler's Third Law

PLANET	DISTANCE	PERIOD	$(D/P)\times 10^3$	$(D^2/P)\times 10^3$	$(D^3/P^2)\times 10^6$
MERCURY	0.387143	87.9583	4.40144	1.7040	7.50
VENUS	0.722467	224.232	3.22197	2.3278	7.50
EARTH	1.000000	365.256	2.73781	2.7378	7.50
MARS	1.524881	687.580	2.21776	3.3818	7.50
JUPITER	5.208507	4340.49	1.19998	6.2501	7.50
SATURN	9.543414	10765.27	0.88650	8.4602	7.50
ANY					7.50

Because the latter product also varies inversely with D/P, BACON.4 computes their product, which is D^3/P^2. On examining this product, BACON.4 discovers that it has the same value for all the observations, hence the desired invariant law has been discovered.

In this description of BACON.4's behavior, a number of important details have been elided. In particular, we have said nothing about the order in which pairs of variables are examined, although that order could clearly affect the amount of search that would be required to find the invariant ratio. In fact, the exact order in which pairs are tested turns out to have only a second-order effect on the program's problem-solving ability. BACON.4 is also provided with checking procedures that allow it to avoid infinite loops – such as calculating the product of D/P and P and adding $DP/P = D$ as a new variable. The important elements that account for BACON.4's efficiency in its search for invariants are: (1) the heuristic that allows it to notice the constancy of a particular dependent variable and to recall the definition of that variable in terms of the observables; and (2) the heuristic that leads it to look for pairs of variables that are positively (negatively) correlated, and to compute their ratios (products).[11]

The new variables that BACON.4 introduces by taking products and ratios of observables or previously computed variables are theoretical terms, in the sense in which we have been using that phrase. They are theoretical terms of a simple kind, being immediately definable in terms of observables. It is important for BACON.4's effectiveness that its heuristics operate on defined terms in exactly the same way as they do on observables. The program is completely recursive in this respect also.

Recursion and Generalization. When BACON.4 has discovered an invariant, it can now summarize the whole set of observations from which it derived that invariant in a single new observation, which shows the value of the invariant and omits all of the independent variables that do *not* influence its value (See the last row of Table 1). This new observation is, in fact, the hypothesis or law that has been discovered. BACON.4 has a completely recursive structure. Having found the Level 1 summaries of Level 0 data for a number of different values of one of the independent variables, it can examine these summaries in exactly the same way as it examined the original data. If successful, it may be able to discover a new, Level 2, generalization that summarizes the Level 1 hypotheses. This process can be repeated for an arbitrary number of levels, the hypotheses being generalized at each step.

To see how this works, consider some data obeying the ideal gas law. This law may be stated as $pV/nT = 8.32$, where p is the pressure on a gas, n is the number of moles, T is the temperature, and V is the volume. Suppose BACON.4 is given data showing that when p is 1, n is 1, and T is 300, the value of V is 2496.0. If the first three terms are under the system's control (independent variables),[12] one can think of their values (p, n, and T) as conditions on the value of V, the dependent variable (see Table 2). Now suppose that after gathering additional data by varying p while holding n and T constant, BACON.4 finds that pV is 2496.0 whenever n is 1 and T is 300. This first level description (Boyle's Law) summarizes all zeroeth level observations that have the same conditions (the same n and T), but it can be treated as data in turn. Upon varying T, the program generates other first-level summaries; taken together these lead to the second level summary that pV/T is 8.32 whenever n is 1 (the laws of Boyle and Charles). Continuing in this way, the system arrives at the ideal gas law when the third level of description is reached.

In determining when to generate a new description to summarize a set of lower level descriptions, BACON.4 draws on a generalized version of the traditional inductive inference rule. It simply looks for recurring values of a dependent variable and, when these are found, hypothesizes this constancy, including as conditions upon it the values of independent variables that are common to all cases in which this constant value occurred. The rule may be stated as:

> IF you see a number of descriptions at level L in which the dependent variable D has the same value, V,

TABLE 2
Derivation of the Ideal Gas Law

Level	Conditions	Indep. Var.	Dep. Var.	Law
0	N, T, P		V	$V = K$
1	N, T	P	V	$PV = K$
2	N	T	PV	$PV/T = K$
3		N	PV/T	$PV/NT = K$

THEN create a new description at level $L+1$ in which the value of D is also V, and which has all conditions common to the observed descriptions.

This production may detect constant dependent terms that take either numerical or nominal (symbolic) values. BACON.4 has primitive faculties for ignoring small amounts of noise in data. However, it cannot, in its present form, deal with significant deviations from regularity, nor can it recover from overgeneralizations once they have been made. The conservative strategy of including all common conditions in the statement of each law serves to offset this latter limitation.

Postulating Intrinsic Properties. Earlier, we compared two forms of the law of conservation of momentum. The first, $a_X/a_Y = -K_{XY}$, asserted that, for any pair of blocks, the ratio of their mutually induced accelerations would be invariant. The second, $-a_X/a_Y = M_Y/M_X$, explained this invariance in terms of the ratios of the inertial masses of the two blocks. BACON.4 is able to discover this second, deeper, form of the law by attributing an intrinsic property, mass, to objects and discovering an appropriate value of the mass to associate with each block used in the experiment.

In this experiment, BACON.4 treats pairs of objects O_X and O_Y as the independent variables and their accelerations, a_X and a_Y as the dependent variables. In addition, the program is told that the objects of a pair are interchangeable (i.e., that the same objects can be used for O_X or O_Y. Thus, when an intrinsic property is discovered for O_Y, BACON.4 knows it can associate that property with the same object when it appears as O_X.[13]

Let BACON.4 experiment with five blocks, A, B, C, D, and E. At level 0 of the experiment, where O_X is the block A, BACON.4 will discover that each pair of objects, A, O_Y, has a constant ratio of accelerations, K_{AY}. Suppose that the ratios are $K_{AB} = 1.20$, $K_{AC} = .80$, $K_{AD} = 1.60$, and $K_{AE} = 1.80$. BACON.4 now defines a new variable, M_Y, an intrinsic property associated with the second object in each pair, O_Y, and sets its value equal to the corresponding K. At this point, BACON.4 cannot specify an M value for block A, which has not appeared in the experiment as O_Y, and so it includes only blocks B, C, D, and E in the remaining experiments it carries out. At this point, BACON.4 also defines the term K_{XY}/M_Y, which, by the definition of M_Y as equal to K_{AY}, must always equal 1.0 whenever O_X is A.

BACON.4 now collects data for new values of O_X. When O_X is the block B, it again discovers a constant ratio of accelerations for each value of O_Y: .667, 1.33, and 1.50 for the pairs (BC), (BD), and (BE), respectively. Since these ratios, K_{BY} vary directly wih the stored values of M_Y, BACON.4 finds that the ratio K_{BY}/M_Y is equal to the constant .833.

At Level 2, BACON.4 now varies the first block, and looks at the values K_{XY}/M_Y obtained on the previous level. When the program retrieves the M_X values of each of the blocks and compares them with the previously mentioned ratios, it finds that their product, $K_{XY}M_X/M_Y$, is a constant. Thus, for any pair of blocks, O_X, O_Y, $M_Y/M_X = K_{XY}$, which is the desired law of conservation of momentum.

Finding Common Divisors. BACON.4 has one additional important heuristic beyond those already mentioned: It examines the several values of a variable that are obtained within any given experimental condition, to see if all these values can be expressed as multiples of some common divisor. Suppose, for example, that in a certain experiment measurements were made of a particular property (the dependent variable) of a number of different chemical elements (the independent variable). The observed values of the property were 0.0446, 0.715, and 0.625, for hydrogen, oxygen, and nitrogen, respectively. BACON.4, upon examining these numbers, would discover that the first (0.0446) divides the others with the integer quotients: 16.0 and 14.0. (The reader may recognize these latter numbers as the atomic weights of oxygen and nitrogen relative to $H = 1$, and the regularity expressed as Prout's hypothesis that all atomic weights are integral multiples of the atomic weight of hydrogen.)

In another example, we might have as independent variable some chemical compounds of oxygen, and as dependent variable, the weight of the oxygen in a standard volume of the compound: nitrous oxide, 0.715; sulfur dioxide, 1.430; and carbon dioxide, 1.430. BACON.4 would notice that the second and third numbers are exactly twice the first number. It was on the basis of this kind of observation that Cannizzaro, following Gay–Lussac and Avogadro, postulated the chemical formulas for these three compounds: N_2O, SO_2, and CO_2, the number of oxygen atoms in each molecule corresponding to the multiple in the weight/volume ratio.

Summary: BACON's Heuristics. We have come to the end of the list of principal heuristics that BACON.4 uses to induce laws from data. The list is very short, consisting of only four items:

1. It detects covariances of variables, and creates new functions of these variables in a search for invariants, carrying out this process recursively with the new variables it has defined.
2. It recurses indefinitely to higher levels, generalizing by varying the independent variables that were held constant at lower levels.
3. It postulates intrinsic properties associated with nominal variables and tests for constancy of these properties.
4. It searches for common divisors among the values of a variable and takes the least common denominator as unit of measurement.

The brevity of the list of heuristics and their independence of the subject matter described by the data show that BACON.4 is itself a rather general and parsimonious theory of the process of law discovery by induction from data.

BACON.4's Performance

To give a more complete picture of the power of BACON.4, as well as its limitations, we summarize here the results of the main experiments we have conducted with it and the laws it has discovered.

In the previous section, we have already outlined how the program can discover Kepler's Third Law, the law of conservation of momentum, and the ideal gas laws. To these, we may add the following:

• *Ohm's Law: Current = Voltage/Resistance*
Given control over two nominal variables, the battery and the wire, and measurements of the current, BACON.4 defines conductance (the reciprocal of resistance) as an intrinsic property of the wires, and voltage (the ratio of current to conductance) as an intrinsic property

of the batteries. It then discovers that current is equal to the product of voltage and conductance, an alternative form of Ohm's Law.

Given, instead, numeric information about the wires (their length and diameter), BACON.4 finds the extension of Ohm's law that gives the conductance (or resistance) as a function of the length and diameter of the wire. In this form of the experiment, the program can also discover the internal resistance of the battery as a second parameter, along with its voltage.

• *Archimedes' Law*
Given two nominal variables, the name of an object and its composition (gold, silver, etc.), and two numeric variables, the volumes of a liquid and combined volumes of object and liquid, BACON.4 determines the volume of each object. Given several objects of each composition, it then proceeds to define the density of each object as an intrinsic property of its composition.

• *Snell's law of refraction*: $\textit{sine}\ i/\textit{sine}\ r = n_1/n_2$
The law relates the angle of incidence and the angle of refraction of a ray of light as it passes from one medium to another. The data given BACON.4 are the sines of the angles. BACON.4 introduces as intrinsic properties the indices of refraction of the media, n_1 and n_2.

• *Black's specific heat law*: $c_1m_1t_1 + c_2m_2t_2 = (c_1m_1 + c_2m_2)t_f$
This law relates the temperatures of two liquids along with their masses to the final temperature, t_f of the mixture. In this experiment the masses were provided as independent variables, while BACON.4 introduced the specific heats, c_1 and c_2, as intrinsic properties of the nominally identified liquids.

• *Laws of chemistry*
BACON.4 has induced a number of the laws of Nineteenth Century chemistry from data on chemical reactions. In examining these reactions, the program treats three variables as independent: the element contributing to the reaction, the resulting compound, and the weight of the element used, w_e. Three dependent variables are measured: the weight of the compound resulting from the reaction, w_c, the (gaseous) volume of the element used, v_e, and the (gaseous) volume of the compound, v_c, under standard conditions.

At the first level, the system finds three constant ratios for each element/compound combination: w_e/w_c, w_e/v_e, and w_e/v_c. The first of these is the relative combining weight of the element, which is constant by Proust's Law. The second is the constant density of the element, and the third is the weight of the element per unit volume of the compound.

At the second level, BACON.4 varies the element and discovers that the w_e/v_c ratios occur in small integer multiples of each other. Using the common divisors discovered in this way as intrinsic properties of the elements, the program goes on to find a common divisor among these latter numbers, thus obtaining the atomic weights of the elements as multiples of the weight of hydrogen.

Finally, BACON.4 discovers that the ratio w_e/v_e depends on the element being considered, but is independent of the compound – equivalent to the statement that each element has a characteristic (gaseous) density.

Thus, BACON.4, without using an atomic hypothesis, obtained the main generalizations arrived at by Proust (law of definite proportions) and Avogadro, and posited the intrinsic properties of gaseous density and atomic and molecular weight. In a related experiment, with a slight re-ordering of independent and dependent variables, the program can discover Gay-Lussac's law (the law of simple proportions by volume).

Does Bacon Explain?

We may now return to our earlier discussion of the relation of explanation to description and ask whether BACON.4 has explained any phenomena, or whether it has only described them. In every one of our experiments, BACON.4 introduced new theoretical terms, hence met at least that criterion of explanation. In most cases, it has also discovered and introduced intrinsic properties, including inertial mass, voltage, resistance (or conductance), internal resistance, specific gravity, index of refraction, specific heat, atomic weight, and molecular weight. Hence, if the use of intrinsic properties is the touchstone of explanation, most of the theories BACON.4 has discovered are explanatory.

What about the test of reduction? The chemical formula, $2H_2 + O_2 = 2H_2O$, provides a reductionist explanation of the formation of water vapor from hydrogen and oxygen, in terms of the atomic theory and hypotheses about the atomic compositions of the respective substances. BACON.4 using its GCD heuristic, discovers that water contains two 'quanta' of hydrogen and one of oxygen, and that each 'quantum' of the latter weighs sixteen times as much as a 'quantum' of the former. Does this result support the claim that BACON.4 has discovered the atomic theory of the formation of water vapor, hence has reductionist capabilities? We would say that it does not, although it is not easy to specify what is missing.

Perhaps we should only assert that a system has an atomic theory if it has some internal representation for atoms, each having associated with it its atomic weight and such other properties as theory attributes to it. Sets of such atoms would represent molecules. In addition to this representational capability, we might require that the system also have operators (representations of reactions) for rearranging sets of atoms.

With such capabilities, the system could simulate the formation of water from hydrogen and oxygen by taking two molecules consisting of two atoms of hydrogen each and one molecule consisting of two atoms of oxygen and rearranging these into two molecules of water. This simulation would then constitute an explanation (or would it be only a description?) of the reaction. If only rearrangement operations were admitted (no creation or annihilation of atoms), the system would obey the basic laws of chemical reactions – conservation of weight, for example, and of numbers of atoms.

It would be no great difficulty for BACON, or a very similar system to discover the conservation of mass and of atoms in chemical reactions (if appropriate data were supplied).[14] Would we then say that BACON held an atomic theory? These are puzzling questions for which it is not easy to find answers. Our own belief is that the answers will be found mainly in the course of trying to construct experimental systems like BACON. Therefore, instead of conjecturing further about these difficult matters, we will describe briefly another discovery system, D. B. Lenat's AM, which illustrates aspects of the discovery process that are not prominent in BACON.4.

DISCOVERY OF MATHEMATICAL CONCEPTS

The goal that the AM system strives to attain is a vague one: to find interesting new concepts and conjectures.[15] It has no machinery for proving or disproving its conjectures 'New ' means 'new to AM,' since the system has no way of knowing what concepts are already abroad in the world.

In contrast to BACON.4, AM is a theory-driven system. Its search for new concepts is guided by the concepts it already has and what it knows about them. However, since AM's domain is mathematics rather than empirical science, there is no way to test its 'theories' – concepts and conjectures – against data. The closest analogue to data in AM are *examples* of concepts. AM does have capabilities for generating examples of its concepts and uses these examples as one basis for guiding the search for new concepts. If the analogy between examples and data is accepted, then to that extent, AM may be regarded as both theory driven and data driven. In this interpretation it exhibits the whole cycle of interaction of theory and data, with examples motivating the construction of new concepts, and concepts motivating the generation of examples.

The Structure of AM

To understand how AM operates and what it can (in principle) do, we must say something about the way it represents concepts in memory, the processes it uses to search for new concepts, and the criteria it uses to guide its search. We begin with its representational scheme.

Representation of Concepts. Each concept in AM is represented as a schema consisting of a list of attribute-value pairs. There is a fixed set of attributes, or 'slots,' which are used to describe all concepts. When a concept is initially introduced, any or all of the slots may be empty, to be filled by further exploration. The 'values' that fill the slots are data structures or procedures of arbitrary complexity.

The standard slots include the concept name, definitions (procedures for generating or recognizing examples of the concept), specializations, generalizations, examples (subclassified as 'typical,' 'barely,' 'not-quite', and so on), conjectures, intuitions (analogic representations), views (mappings on to other concepts), analogies, algorithms (for executing the concept on a given argument), domain, range, worth, interestingness, and a few others. We will not explain in detail the content of each of these slots. We list them to indicate the degree of 'richness' a concept can acquire, and of interrelation with other concepts.

A new concept is created by creating a node in memory, and assigning to it a name and at least one definition. How far it will later be elaborated by the slot-filling processes will depend on its ratings for worth and interestingness.

Initial information. At the start, AM is provided with three packages of information:

1. Criteria to measure how interesting a concept or conjecture is. For instance, AM judges a concept to be interesting to the extent that it is related to other interesting concepts, to the extent that examples of it can be generated (without too much difficulty but not too easily), and to the extent that strong conjectures can be made with its help. An example of a concept is interesting to the extent that it is extreme (is barely within the boundaries of the concept). The interest criteria, of which there are a substantial number, are assigned real numbers that are added together to determine the net worth of a concept or conjecture.

2. Heuristics for searching for interesting concepts. For instance, AM searches for new concepts by constructing examples. If examples of a concept are very easy to find, it specializes the concept; if they are very hard to find, it generalizes the concept. AM also specializes concepts to create narrower concepts that fit extreme examples of the original concept. It has combinatory capabilities for generating new concepts by assembling tests from concepts already in its stock. Given an interesting concept, AM may construct its inverse, and other concepts related to it.

3. Finally, AM is provided with an initial stock of concepts in some subject-matter domain.

The two principal experiments with AM used the domains of elementary geometry and set theory. In set theory, AM was given the definition of set, of union of sets, of intersection of sets, and so on – a total of some 115 concepts. This was the kind of information that would be introduced in a first course on set theory.

AM is not a parsimonious system. We have noted that in its exploration into set theory it was supplied with more than one hundred initial concepts. The criteria for evaluating interest and the heuristics for searching for interesting concepts and conjectures also number in the hundreds. Most of the criteria and heuristics correspond to fairly general weak methods, applicable to many different mathematical domains; a few correspond to more powerful, domain specific, methods. There is no objective standard for judging how closely they are tailored to the particular concepts that AM discovers, but our conclusion, after examining the system closely, is that they cannot generally be accused of being ad hoc.

Search Organization. The control system of AM directs a best-first search. At each episode, AM selects from an agenda list the task that has the highest worth, and allocates a quantum of computing power to it. When the allocation of computing resources has been exhausted, AM makes a new evaluation of the available tasks and resumes search with the most promising. The product of task activity is to define new concepts, fill slots of existing concepts, and add new tasks to the agenda after evaluating them. Hence, the system has no goal more definite than elaborating concepts and searching for new concepts in the neighborhood of the most interesting things it knows already.

Performance of AM

In the domain of set theory, after using about one hour of central processor time on a PDP 10Kl, AM had discovered the natural numbers and the basic operations of addition, subtraction, multiplication, and division upon them. It had assigned a high interest value to the operation of division, and had examined examples of numbers with different numbers of factors. It had paid particular attention to numbers possessing only two factors (i.e., the prime numbers). It had conjectured that any number can be represented uniquely by a product of powers of primes (the so-called 'fundamental theorem of arithmetic') and that any even number can be represented as a sum of two primes (Goldbach's conjecture, not yet proved or disproved).

AM then began to make conjectures about numbers with maximal numbers of primes (for their size). Here it appeared that the program, for the first time, was on completely new ground beyond the present literature of mathematics. Alas, it was not so. Ramanujan, the self-taught Indian mathematician, had once gone that way.

CONCLUSION

What light do BACON.4 and AM throw on the hypothesis that the processes of scientific discovery are simply the processes of problem solving, familiar to us from many other problem solving domains? What can we conclude from our examination of these two systems?

First, some limitations. Several aspects of scientific discovery lie outside the scope of both programs.[16] BACON and AM do not gather data by making observations or conducting experiments. They do not design experiments (but see the last footnote, with reference to the MOLGEN program). They do not invent new instruments or methods of observation.[17] Unless we accept the analogy, proposed earlier, between AM's examples and empirical data, the programs do not make theory-driven discoveries. Finally, they do not, in any clear way, accomplish what we would recognize as explanation by reduction.

Neither AM nor BACON.4 can claim to have made a discovery that was unknown to the world of science, though each system made, as we have seen, numerous discoveries that were new to *it*. Until at least one substantial wholly new discovery has been made by such a system, it may be claimed that, implicitly at least, the requisite knowledge was given to it by its programmers. We know no formal procedure for evaluating a claim of this kind. All one can do is to examine the program as carefully as possible for evidence of (explicit or implicit) ad hoc assumptions. In our judgment, both AM and BACON are reasonably free from such evidences.

On the positive side, it is not very difficult to evaluate the quality of the discoveries made by AM and BACON. In both cases we can look to the histories of science and mathematics for an assessment. Both AM and BACON.4 have rediscovered concepts or laws that were considered to be of first magnitude by the contemporaries of the original discoverers. Taken collectively, BACON and AM do model a range of scientific discovery activities – if not the whole gamut – and perform impressively within that range.

In their basic organization, both AM and BACON are examples of heuristic search systems of a kind that is wholly familiar among artificial intelligence problem solvers. Of course, the search spaces and heuristics are adapted to their domains of inquiry, but are nevertheless quite general.

On the basis of the experience with these programs, it seems reasonable to claim that the mechanisms of scientific discovery can, indeed, be subsumed as special cases of the general mechanisms of human problem solving. To be sure, there may be essential novelty hidden in those aspects of problem solving that lie outside the range of the programs. But given the evidence of behavior that we have reviewed here, bare claims that such novelties exist are not convincing. There seems to be no present reason to believe that any aspects of scientific discovery must remain indefinitely beyond the powers of heuristic search, or that the discoveries of human scientists cannot in time be explained within the information processing paradigm for problem solving.

Carnegie-Mellon University

NOTES

* This research was supported by Grants SPI-7914852 and IST-7918266 from the National Science Foundation, Research Grant MH-07722 from the National Institute of Mental Health and Grant F33615-78-C-1551 from the Advance Research Projects Agency of the U.S. Government.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of any of the sponsoring agencies or the U.S. Government. File Disc. Mss.

[1] *See H. A. Simon, Models of Discovery* (Dordrecht: D. Reidel, 1977), Chapter 5.2, 'Scientific discovery and the psychology of problem solving,' and Chapter 5.4, 'Does scientific discovery have a logic?' , and A. Newell, J. C. Shaw, and H. A. Simon, 'The processes of creative thinking,' reprinted as Chapter 4.1 in H. A. Simon, *Models of Thought* (New Haven: Yale University Press, 1979).

[2] See *Models of Thought*, op.cit., Section 5.

[3] D. B. Lenat, 'Automated theory formation in mathematics,' *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 1977, pp. 833–842.

[4] P. W. Langley, 'Rediscovering physics with BACON.3,' *Proceedings of the Sixth International Joint Conference on Artificial Intelligence*, 1979, pp. 505–507; G. L. Bradshaw, P. W. Langley, and H. A. Simon, 'BACON.4: The discovery of intrinsic properties,' *Proceedings of the Third National Conference of the Canadian Society for Computational Studies of Intelligence*, 1980, pp. 19–25; and P. W. Langley, G. L. Bradshaw, and H. A. Simon, 'Rediscovering chemistry with BACON.4,' forthcoming.

[5] For a defense of the possibility of such a normative theory, see *Models of Discovery*, op. cit., Chapter 5.4.
[6] For a general view of this contrast in problem solving by experts and novices, respectively, see J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, 'Expert and novice performance in solving physics problems,' *Science* **208**, 1335–1342 (June 20, 1980).
[7] See A. Newell and H. A. Simon, *Human Problem Solving* (Englewood Cliffs, N. J.: Prentice-Hall, 1972) and Nils Nilsson, *Principles of Artificial Intelligence*, (Palo Alto, Cal.: Tioga Publishing Co., 1980).
[8] On Mendeleev see B. M. Kedrov, 'On the question of the psychology of scientific creativity,' *Soviet Psychology* **5**, 18–37 (Winter 1966–67). An account of Balmer's discovery may be found in L. Banet, 'Evolution of the Balmer Series,' *American Journal of Physics* **34** (1966), 496–503.
[9] For a thorough discussion of the notion of 'theoretical term' and references to the literature on the topic see R. Tuomela, *Theoretical Concepts* (New York: Springer-Verlag, 1973).
[10] See *Models of Discovery*, op cit., Section 2, 'Causes and possible worlds.'
[11] More precisely, BACON.4 looks first for linear relations between pairs of variables, computing the ratios or products when one is not found.
[12] In carrying out this experiment, we treated V as the dependent variable and p as an independent variable. In the laboratory, we would usually control the former rather than the latter. However, we describe the experiment here as it was actually given to BACON.4. The reversal of variables does not affect the result.
[13] The present discussion describes an early version of the procedure for defining intrinsic properties. For a description of the more conservative approach now taken, see Langley, Bradshaw, and Simon, 'Rediscovering chemistry with BACON.4,' *forthcoming*. The earlier version is used here because it makes for a simple exposition.
[14] Milikan used the procedure of finding a common divisor to induce the quantified charge of the electron and the atomicity of that particle from the data of his oil-drop experiment.
[15] The description of AM in this section is based on D. B. Lenat, op. cit. Of course, the present authors are responsible for our interpretive comments about the system, which should not be attributed to its designer.
[16] However, note that we have not, in this paper, surveyed the whole population of programs that are capable of making scientific discoveries. In addition to AM and BACON.4, there are at least HEURISTIC DENDRAL, which, like BACON, induces theories from data, and MOLGEN, which designs experiments in molecular genetics. For the former, see B. G. Buchanan, G. Sutherland and E. A. Feigenbaum, 'Heuristic Dendral: a program for generating explanatory processes in organic chemistry,' in Meltzer and Michie, eds., *Machine Intelligence* **4** (New York: American Elsevier, 1969), pp. 209–54. For the latter, see N. Martin, P. Friedland, J. King, and M. Stefik, 'Knowledge base management for experiment planning in molecular genetics,' *Proceedings of the Fifth International Joint Conference on Artificial Intelligence*, 1977, pp. 882–887.
[17] However, in its discovery of Archimedes' Law, BACON.4 invents, in a sense, a method for measuring the volumes of irregular objects.