INTERNATIONAL STUDIES IN THE PHILOSOPHY OF SCIENCE, VOL. 6, NO. 1, 1992

Scientific discovery as problem solving: reply to critics

HERBERT A. SIMON
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Since the papers discussing my theory of scientific discovery focus mainly on the same few issues, to avoid redundancy in my reply I will deal first with some of these issues of common concern and then comment briefly on other points made by individual commentators.

The fallacy of the definite article

(See also comments on Gillies, Losee, Marconi, Schank & Hughes, and Watkins.)

A number of the critics commit the Fallacy of the Definite Article. To use the phrase 'the X' implies (1) that the thing denoted X exists, and (2) that it is unique. Several of the commentators argue that my theory of scientific discovery does not deal with "the real" essence of scientific discovery', but only with some of its relatively inessential concomitants. In employing the definite article, sometimes emphasized by adjective and noun, these critics are asserting that 'real' creativity is to be found in only one of the many activities in which scientists engage. They are not in complete agreement as to what this essence is, but they seem to favour two candidates: (a) finding the research problem, or (b) finding a good representation for it.

Now my paper, and the longer discussion of the same issue in Chapters 1 and 10 of *Scientific Discovery*, with Langley *et al.* (1987) show that there is no such thing as "the creative step in the discovery process". As I stated in my paper, scientists do many things, and the progress of science depends on all of them being done, each contributing to the possibility and success of the others. Scientists sometimes find new problems and new problem representations; but they also sometimes make observations, perform experiments, observe new phenomena, invent new instruments, induce new laws from data and/or theoretical premises, induce explanatory laws from descriptive ones or descriptive predictions from laws, and plan experimental strategies.

Not all scientists do all of these things; some do just one or two. The degree of specialization differs from science to science. In our research on scientific discovery and creativity, we have applied a sociological and historical test to determine which steps in the scientific process embody 'the real creativity'. We have simply asked, "What steps in scientific discovery win for the discoverer acclaim as a scientist?" Put more crudely, "What does it take to get your picture in a book on the history of science, or to win a Nobel Prize?"

The verdict of history is that taking any one of these steps may constitute important creative science. You can win prizes and a place in the history books by finding a new

scientific problem (but rather rarely), by finding a new representation for a difficult problem or even a new kind of problem representation (e.g. by inventing the calculus), by solving a problem that has already been stated and represented, or by doing any one of the other things listed above. It is easy to find examples of Nobel prizes awarded for each of these activities, except possibly for finding new problems (without solving them). Hence, the claim that 'the real discovery is X' commits the Fallacy of the Definite Article.

I will now develop this point further with respect to finding new problems and finding representations. For each, there is both a negative and a positive point to be made. The negative point is that in many celebrated cases of discovery, the discoverer neither set the problem, which was already well known, nor provided a new representation for it. So if solving problems without contributing to these previous two steps is not creative, we will have to cut out a lot of portaits from the history books.

The positive point is that the processes of finding problems and of finding representations are themselves problem solving processes. Although these processes have not been explored as thoroughly as some other discovery processes, enough is known about them so that they are not mysterious.

Let me proceed with the argument in more detail.

Finding problems

(See also comments on Gillies, Marconi, and Schank & Hughes.)

As for the negative point, Kepler did not invent the problem that he solved with his Third Law. That problem was first proposed by Aristotle, who mentions that the more distant planets revolve more slowly than the nearer ones. The data that Kepler used were not original either, but were obtained from Copernicus. (The elliptical shapes of the orbits were not relevant; only their average distances from the Sun.) Nor was the choice of variables a problem. How many properties of the planets had been measured? Period of revolution, colour and brightness, and finally distance. Neither Kepler nor BACON had any trouble in choosing variables to relate.

Newton was not the sole inventor of the problem of determining the law of gravitation; the problem had occurred to Hooke, Wren, and others—and they had even conjectured the inverse-square form of the answer. Newton's contribution was to show rigorously that this answer predicted the data, for example the Moon's orbit, correctly.

Similarly, Joseph Black did not formulate the problem of finding the equilibrium temperature of a mixture of liquids. He solved it. It had been studied a generation earlier by Farenheit and Boerhaave, who used exactly the same experimental paradigm as did Black, but arrived at the wrong law. It is Black who (quite properly) receives the acclaim, not his distinguished predecessors.

Einstein was not the first to pose the problem he solved with the theory of special relativity. All of the anomalies that created the problem were well known to his contemporaries, and at least two distinguished physicists (Lorentz and Fitzgerald) were within a hairbreadth of the solution (had, in fact, found the Lorentz transformation). Nor did Einstein invent the problem of the photoelectric effect, whose solution won him the Nobel Prize.

Planck did not invent the problem of finding the law of blackbody radiation; it had been posed by Helmholtz some 30 years before the young Planck went to work on it. Nor did Planck generate the data. Having proposed the law in 1900, and having provided an explanation for it the same year, it was not until 1905 that its real meaning and the significance of the discrete quantum were explicated by Einstein and Ehrenfurst; and it

was some years later that Planck understood it. Shall we remove Planck from the history books?

Do these examples suffice? If not, there are many more. The jury rejects the claim that finding the problem, rather than solving it, is where 'the real creativity' resides.

However, to deny that finding the problem is the really creative act is not to deny that finding a problem may be creative. How is it done? Is there any evidence that problems are found by problem-solving processes? One good way to find an important problem is to search among phenomena for one that is important or interesting. If you search at random, you will likely be a long time in finding anything significant. Therefore, you need to search selectively, using whatever heuristics you can lay hands on.

One heuristic is to be alert for surprising phenomena. To be surprised, you must have expectations that are violated by the surprise; and to have expectations, you must have knowledge of what to expect. Kulkarni's program, KEKADA, forms expectations about what outcomes to expect from its experiments. When it is surprised by the disappointment of one of these expectations, its next problem is to explain the surprising phenomenon.

When ornithine produced an unexpectedly large yield of urea, KEKADA (like Krebs earlier) took up the problem of explaining this phenomenon. When Fleming was surprised to find a mold lysing the bacteria in his Petri dish, he sought out the cause of this effect. When the Curies noticed a source of radiation more intense than that of the uranium they were refining, they persisted until they discovered radium.

Another heuristic for problem selection is to apply a novel technique. Krebs learned from Warburg how to experiment with tissue slices. What should he do with his new scientific weapon? He looked around for recognized problems that had not been solved by methods previously available, and found the problem of urea synthesis. A new instrument can also define good scientific problems. The availability of the thermometer set off a whole host of experiments on temperature; the availability of the ammeter, a whole host of experiments on electricity.

There is no mystery or magic in problem generation. Some problems come from surprising phenomena, and these phenomena present themselves to the prepared mind, often, but not always, as a result of using new experimental paradigms or new instruments. However, there are many other ways in which finding problems can become an exercise in problem solving. Pages 302–312 of *Scientific Discovery* discuss a whole host of heuristics for finding problems, providing further support for the hypothesis that problem finding is a form of problem solving.

Finding representations

(See also comments on Cordeschi, De Mey, Johnson-Laird, Losee, Marconi and Watkins.)

Finding representations is also not *the* creative step in scientific discovery. It is not because many scientists are judged to be highly creative for solving problems using representations already available to them. Clark Maxwell did not invent a new representation in order to state his laws of electromagnetism. Nor did he invent the problem, which goes back to Faraday and beyond.

One of the first representations Maxwell employed derived from a hydraulic analogy proposed by Faraday. Maxwell then turned to partial differential equations, which were already there to be used; and, in fact, had been used by William Thomson (Lord Kelvin) to attempt his earlier theory of electromagnetism. It was Maxwell, who found the right equations. The highest recognition went to the scientist who solved the problem, not

the one who first proposed the fruitful problem representation. I am wholly prepared to honour both. Both steps were essential to solving the problem; neither one was the creative act.

Kepler proposed no wholly new representation for astronomical phenomena. The general scheme—of relating the positions of the planets to those of the fixed stars—goes back to Ptolemy and beyond; its heliocentric interpretation to Copernicus. The notion of elliptical orbits, if we want to call that a separate representation, was Kepler's own, but it played no role in the discovery of his Third Law.

Bohr did not invent the analogy between an atom and a planetary system. This had been proposed by Rutherford, but ran into trouble because the magnetic field of the moving electrons would dissipate energy, quickly winding the system down. Bohr took over the planetary metaphor, but denied the working of the laws of electromagnetics in this context, and replaced them with the quantized energies of Planck (or Einstein or Ehrenfurst), and the descriptive formula of Balmer for the hydrogen spectrum. To whom shall we give the laurel? Was Bohr a creative scientist?

The idea that DNA strands were helical was not invented by Crick and Watson. Helical structures had been observed in proteins and were known to Pauling (and many others). To be sure, the double helix was a representational innovation, but an innovation obtained by modifying existing representations in the face of detailed data that provided essential constraints.

Perhaps this is enough for the negative case: to demonstrate that inventing representations is not the creative step. This does not deny that such invention may be an important and creative component in scientific discovery. Can we say anything about how it is done? On pages 315–326 of *Scientific Discovery* the topic is discussed at some length. See especially the description, on pages 317–318, of the computer program, UNDERSTAND, which reads problems in English prose and constructs from them representations (sets of list structures) that are sufficient to allow the General Problem Solver to go to work solving the problems. However, we can go further.

Finding the right representation for a problem usually means drawing on the modest stock of representations that is already known. Such new representations as the differential calculus are of extreme rarity, and not many creative scientists can claim ever to have invented one. Crick and Watson drew on quite standard representations for chemical structures, including helical ones. They gradually induced the correct one by attending to an array of crystalographic and other evidence that ruled out alternatives. Watson himself makes no claim that other scientists (e.g. Pauling) would not soon have solved the problem if they had had access to the same data. A jigsaw puzzle becomes solvable when the pieces are there.

Recently, Craig Kaplan and I have found evidence indicating how the right representation is found for a celebrated AI problem whose solution depends on changing the representation: the Mutilated Checkerboard problem. Consider a checkerboard with one-inch squares, and a set of 32 dominoes, each 1 by 2 inches. Clearly, we can cover the entire checkerboard with the 32 dominoes, each covering two squares. Now suppose that we remove the north-west and south-east squares of the checkerboard, so that 62 squares remain. Can we cover all of these with 31 dominoes?

Subjects generally try various possible coverings, of course, without success. Typically, they persist for an hour or two. Only when they are severely frustrated do they consider using a different representation, but at first they are unable to generate one. Apparently, our species does not come equipped with a general representation generator.

Some subjects notice eventually that the squares they fail to cover are always of the same colour. Once they note that, they quickly observe that each domino covers one square of each colour, but that the two deleted squares were of the same colour. Hence, the mutilated checkerboard has more squares of one colour than of the other, but the dominoes can only cover the same number of squares of each colour. These insights come rather rapidly once the subjects focus attention on the relevant variables and abstract from the checkerboard to consider only the numbers of squares of each colour that are covered.

This focusing of attention is caused, at least in part, by the invariance, over repeated trials, of the colour of the uncovered squares. Kaplan has written a computer program that can notice this invariance and create the new abstracted representation. Hence, we have shown, at least for this specific (but difficult) problem, how change in representation can be achieved by normal problem solving processes.

Although much remains to be learned about how problem representations are formed or discovered, perhaps enough progress has been made on this subject to shift the burden of proof to the sceptics. We can now point to several processes that solve the problem of finding representations. These processes resemble other problem solving processes already familiar to us—for example, processes for analogizing, and for abstracting and planning. They support the hypothesis that finding problem representations is problem solving.

Intuition again

(See also comments on Agassi and Johnson-Laird.)

The three I's, Inspiration, Insight and Intuition, seem to generate perpetual wonder. The only way to return them from the world of the miraculous to the world of the real is to require some operational tests that signal the phenomena more or less unequivocally. If we can specifiy just when inspiration, insight or intuition has occurred, then we can investigate the causal factors and processes that underlie them.

Of course, vitalists might deny that such explanations are possible. To that objection I have no answer, but will address myself to those readers who suppose that the phenomena of human thinking can be accounted for by natural mechanisms residing in the human nervous system.

We say that someone has an insight or an intuition (or even inspiration) when he or she answers a question or solves a problem rather rapidly (sometimes we say 'instantaneously', but that is poetic license; the time is never less than a second and usually much more), and especially if the respondent is not able to explain in any detail how the answer or solution was reached.

When pressed for information about the method of solution, the respondent may reply, "I just used my intuition", or "it's based on my experience", or "the answer just came to me". There is no reason to doubt the truth of these replies. They are just what we could expect if the solution were obtained by an act of recognition: that is, if some cue in the stimulus situation evoked a recognition of something already familiar in the mind of the respondent, and thereby gave access to information previously stored in memory.

In my paper I show that recognition is a well-understood process in psychology that has been simulated effectively by the EPAM program (Feigenbaum & Simon, 1984). It is also well known that a person can report what he or she has recognized, but not what features of the stimulus allowed it to be discriminated from other possible stimuli. The discrimination process is subconscious, hence not reportable. Recognition is 'intuitive', or better, intuition is simply recognition.

It follows that people are likely to have valid intuitions only about matters in which they are more or less knowledgeable and experienced. For example, a chess grandmaster, in an exhibition of simultaneous play against many opponents, can win by intuitive play (at least against experts and even masters). This calls for almost no forward analysis, but depends upon the recognition of errors that the weaker opponent makes in the choice of moves, errors that leave behind telltale cues ('double pawns', a 'weak bishop', an 'open file' and so on). Similarly, an experienced physician can often make an 'intuitive' diagnosis (of course, usually to be checked by obtaining additional information and performing tests) on the basis of superficial symptoms, visible or reported by the patient.

So competent scientists do often make discoveries by 'intuition'. That is to say, they notice phenomena that others, less knowledgeable, would not notice; they have heuristics for responding to such phenomena; they have much stored knowledge about the implications of the phenomena; and so on. Thus Fleming, noticing that some bacteria are being lysed in the vicinity of a growth of mold on the Petri dish, is surprised and takes steps to determine the scope of the phenomena (what kinds of bacteria?) and its mechanism (what substance is the mold excreting?). These recognition processes are implemented in such simulations as DALTON and KEKADA, and are an important component of the competence of these programs. Hence, intuitive, insightful and inspirational processes are normal problem solving processes that do not introduce any new elements into the theory of scientific discovery.

The incremental character of discovery

(See also comments on de Mey, Hesse, Johnson-Laird, Losee, and Watkins.)

One source of the Fallacy of the Definite Article is failure to remember that science is an incremental activity, each dwarfish act standing, as Newton's metaphor put it, on the shoulders of the giant ones that preceded it. What causes the Moon to be where is is right now? Well, of course, the forces of gravity together with the Moon's position a moment ago. However, what caused that previous position? The forces of gravity again, and the position before that.

In my paper, I posited that a theory of scientific discovery is a set of laws that shows how each new step of progress arose from the previous step. If the laws of mechanics are expressed in differential equations to account for the path of a system through continuous time, the laws of scientific discovery are expressed in computer programs, which have the formal character of systems of difference equations, and account for the progress through successive discrete steps in time.

It is no valid objection to the explanation of a discovery that the explanation refers back to a set of initial conditions that are not themselves explained. We may explain, for example, how Kepler matched his law to the periods and distances of the planets without being held to account for the sources of his data of periods and distances or of his motivation for dealing with just this problem.

Of course, we will want our theory to explain these antecedent events as well. Hence, as it advances, it must become a theory of finding problems, of finding representations, of finding data, of inventing instruments, as well as a theory of finding laws to fit data or mechanisms to explain these laws. In my paper and in my responses to criticisms I have tried to show how far we have progressed with the various components of this theory.

In time, we will want to specify not only these components, but the control structures that assemble them into a coherent activity. Chapter 9 of *Scientific Discovery*, while it does not provide definitive answers to these questions, addresses them at some length. We may

think of assembling the steps of scientific discovery into the path of science along the following lines.

A step of scientific progress ends when some new knowledge has been stored in the mind of the scientist or communicated to the scientific community through publication. We can think of the mind as a private blackboard for the scientist, and the publication as a public blackboard. Consider the latter. Every scientist can take it as a starting point for his or her next step of scientific activity: as a source of ideas for a new problem, a new representation, new data, new instruments, new theories. We can think of BACON, DALTON, KEKADA and the other discovery programs as specialized scientists who take as their initial conditions the current state of the blackboard (and of their private blackboards).

Using the blackboard, the specialized processes assemble themselves quite automatically, without need for central planning. The activity of science, writ large, is a production system (in the computer science meaning of that term) operating off the blackboard of the literature. The individual scientists, the individual programs, constitute its difference equations, and the content of these equations provides the only theory that we can have, or need to have, of the operation of the system. The theory is incremental. The equations explain how the next step is generated by the state of the system and its boundary conditions.

A system of this kind can accommodate extrinsic as well as intrinsic causes in science. Social processes external to science may place premises on the blackboard that influence the motivation of scientists, the problems they select, their criteria of solution and so on. How important extrinsic influences are, in comparison with intrinsic ones, is an empirical question.

The problem of induction

(See comments on Agassi, Johnson-Laird, and Petroni.)

Can there be a logic of scientific discovery? Again, we must be careful to define our terms. By a 'logic of scientific discovery', I simply mean (and have always meant) a set of normative rules, heuristic in nature, that enhances the success of those who use them (as compared with those who don't) in making scientific discoveries. My dictionary legitimates this usage by defining 'logic' as 'a particular mode of reasoning viewed as valid or faulty', and by speaking of 'the relevance, propriety, or logic of some action'. By a 'logic of scientific discovery' I assuredly do *not* mean a formal system of reasoning that provides guarantees of finding scientific laws, or guarantees that the laws it does find will provide a valid theory of the phenomena. Life, alas, does not provide such guarantees.

When BACON, or any other discovery program, searches for and finds a law that fits some finite set of observations, it has made a scientific discovery, which may, on subsequent examination of other evidence, turn out to be valid or not. Scientists must get their pleasure from the beauty of the patterns they discover, for they can have no final assurance that these patterns will remain valid as new data and theories are introduced.

A discovery program certainly does not solve the classical problem of induction: that is, it does not provide evidence that the law it has discovered, however well it fits the finite body of available data, will have predictive validity. The same thing may be said, word for word, about the laws discovered by human scientists.

So the task of a scientist (or a discovery program) is not to achieve the impossible goal of inducing a valid general law from a finite body of data. Their task is to induce a law that fits the data, with the hope (buoyed up by history and prior experience) that the

generalization thus discovered may retain its validity as new data are obtained and new phenomena observed. That's what Newton did when he proposed his Laws of Motion. In the long run, he turned out to be wrong, as the anomalies that were later removed by special and general relativity and quantum theory demonstrated, but we would not conclude from this long-run outcome that his search procedures lacked logic (relevance and appropriateness).

BACON's successes in fitting laws to finite sets of data demonstrate that the search heuristics it embodies constitute a logical (i.e. relevant, appropriate, but certainly not optimal) method for seeking regularity in nature. There is neither claim nor guarantee that the regularities thus discovered will be extended to expanded bodies of data. The problem of induction remains an unsolved problem—to my mind, a problem that neither can be solved nor, fortunately, needs to be solved in order to carry on science. In this respect, BACON is neither better nor worse off than Kepler, Newton, Joseph Black or Balmer.

The process of finding laws, like other problem solving processes, involves an alternation of generation steps and test steps. The system or scientist generates laws that may fit the data, tests these laws against available data, and if they do not fit, generates new laws, guided by the nature of the mismatches between data and previous hypotheses. As Popper pointed out long ago, the test is always a tentative one; we can fail it, but we cannot permanently pass it—new data may always refute our hypothesis.

Now there are silly (i.e. illogical) ways to carry on such a search. Generating laws at random is one of them. There are also reasonable, relevant, appropriate ways to do it, by making the direction of search sensitive to what is found along the way. Characterizing these more effective ways of search is what the logic of scientific discovery is all about. Discovery processes may be 'logical' in the same sense that processes of medical diagnosis or engineering design may be 'logical', and in no stronger sense. The doubts that have been expressed about the possibility of a logic of scientific discovery refer to some other meaning of the word 'logic' than the one I have consistently employed.

Comments of individuals

I turn now to objections raised by individual critics, not already dealt with adequately under the previous topic headings.

Joseph Agassi

Professor Agassi objects to my "exaggerated claim" that our programs "sanction the replacement of live human researchers with computers", and asserts that I claim "for these programs the ability to discover any possible scientific law". The programs themselves show to what extent they can do various kinds of things that human scientists do; I have tried to indicate above, and in all my other writing, some of the things they cannot at present do. If this is "sanctioning the replacement of humans", so be it. Professor Agassi's objection to my "exaggerated claim" is itself an exaggeration of that claim.

As far as "any possible scientific law" is concerned, I cannot conceive of what that phrase means or where this alleged claim was stated. Hence, Professor Agassi is, as he fears, erroneous in his attribution of these claims and I hope that I have been clear about them here.

No program can now do what Bohr did. This indicates an incompleteness (one of many) in the current theory of scientific discovery. It does not show that the theory is wrong about the phenomena it covers. Incompleteness is not refutation. The fact that quantum mechanics provides today only an unsatisfactory and incomplete theory of superconductivity is not a disproof of quantum mechanics, nor are the failures of meteorological

prediction a disproof of the laws of aerodynamics. No physical theory exists today that can handle computationally more than a tiny fraction of the phenomena it purports to explain. In astronomy, not even the three-body problem has been fully solved.

Similarly, the lack today of a simulation of Bohr is not disproof that the processes of scientific discovery are problem solving processes. It is not a condition of the theory that the existing program be "so rich that it can perform all the tasks required of scientific research".

Next, Professor Agassi questions the novelty of my argument. I feel no urge to argue its novelty, just its correctness. Nor do I claim that BACON and the other programs have made discoveries. I have always called them rediscoveries. Their significance lies not in their novelty, but in their demonstration of a set of processes sufficient to make the rediscoveries, starting from the same initial conditions as the original discoverer—hence, providing a tested theory of the discovery—a set of difference equations that accounts for the observed data of discovery.

We have no intention of copyrighting our programs, but copyrighted or not, it is easy to see how they could replace live researches. Whether such researches are 'drudgery' or fun has to be judged by the individual researchers. The affect revealed in such terms as 'drudgery' and 'science-making sausage-machine' is understandable, a quite common human reaction, but hardly relevant to the question before us.

Is it true that "the involvement of luck may render any discovery unrepeatable and so not given to scientific study"? Statistical decision theory provides the tools we need for analysing data in the presence of noise (i.e. randomness or luck). Since all data are noisy, Professor Agassiz's proposition, if true, would make all science impossible.

Nor need we have nightmares about "a super-powerful computer that can correlate every possible correlation between all available data". First, there obviously can be no such computer. Secondly, this is not what computer programs for simulating discovery do. They sometimes find laws, after a small amount of very selective computing, by heuristic search, even as you and I.

As to the extension of our theories to revolutionary science, we can already say a great deal about the processes underlying such discoveries as the law of black body radiation (see *Scientific Discovery*, pp. 47–54) or plate techtonics. Are these normal science?

Professor Agassiz thinks our models omit the role of criticism in research. (I will not take up his problems with editors; I sometimes have problems with editors, too.) On the contrary, our models involve both generating theories and testing them—and rejecting them when they fail to fit the data.

Finally, Professor Agassiz's last paragraph appears to be an apotheosis of inspiration. I have already explained how all the phenomena attributed to inspiration can be explained as recognition. I am grateful for Professor Agassiz's good wishes, and concur in his appreciation of the dangers as well as usefulness of all new knowledge.

Roberto Cordeschi

Professor Cordeschi's comments focus, as he says, on problems of representation; and since I have already commented on that topic, I can be brief.

He raises the interesting and important question of how we can tell what has been 'hidden' in the structure of the simulation program itself, and turns to Lenat's AM as an example. Now, of course, a computer can only do what it is programmed to do (even as you and I), and in a certain sense, everything that comes out has been programmed in. However, if that is a fault, it is a fault that infects all mathematics and all correct reasoning.

Mathematics is incorrect if the conclusions are stronger than the premises; if the theorems are not 'hidden' in the axioms.

In empirical science, and programs like BACON, the inputs include not only the program's (or scientist's) knowledge and heuristics, but the empirical data it is trying to describe. The result will be a reflection of those data; the program cannot claim success until the law it finds fits the data.

With respect to AM, Lenat in his paper with Brown perhaps conceded too much and too soon. A subsequent paper in *Artificial Intelligence* by Weimin Shen (1989) demonstrates that AM's discoveries did not, in fact, depend on the structure of LISP, since they can be made at least as easily by a program written in a simple functional language that has almost none of LISP's characteristics.

Professor Cordeschi's final comments on semantics do, indeed, give 'food for thought'. Whether scientific discovery systems use symbols to refer, or whether they 'reflect on their own actions' are interesting issues. The laws BACON finds surely refer to the data they describe. Of course, these data were gathered by someone else, and BACON has no direct contact with their real-world denotations (nor do our eyes have such contact with the objects they 'see'). Perhaps this says something about science, and even about epistemology.

A program, like DALTON, that attempts explanations is another matter. Here the data, volumes and masses of reagents, are re-interpreted in a second representation, one of hypothesized atoms and molecules. True, DALTON does not see these objects, but then, we don't either. The operationally defined entities that connect DALTON to the external world are the volumes and masses. They are interpeted, by DALTON and by us, as manifestations of the behaviours of atoms and molecules. Where is the difference? What, if anything, is lacking in DALTON's semantics?

DALTON does not have real sense organs. However, robots exist today that do and we could connect DALTON to one of those to collect the data. Would DALTON's semantics then be more robust?

Finally, is it true that these programs do not reflect on their own actions? DALTON tries to fit an atomic model to its data. When the model doesn't fit, DALTON notices this and tries to fit another. Is it reflecting on its own actions? Do we have different ways of reflecting on our actions? Does reflection mean more than evaluating the effects of the actions and guiding them accordingly? I will not try to answer these questions here.

Marc De Mey

I am fascinated by Professor De Mey's quotations from Oresme, showing that demystification has a long heritage; and grateful for his elegant summary of our theory. He raises a tough question: If scientific discovery is so easily explained, why is it so difficult to achieve?

We can explain only incompletely why it took Kepler several decades to do what BACON did in two minutes and some of our college-student subjects did in an hour. We do address this problem on pp. 111–114 of *Scientific Discovery*. Kepler devoted only a small fraction of his waking time over those decades (1%?) to the Third Law. There was a hiatus of 10 years after he discovered a (subsequently rejected) approximation, and during this interval he was preoccupied with other matters, such as his mother's trial for witchcraft.

Kepler had to carry on his work using the human nervous system, which is notoriously slow (milliseconds) compared with a computer (microseconds). He did his arithmetic without a calculator and without logarithms (which he only learned about afterwards). He

reports that he made an arithmetic mistake that stretched out his final search from no more than several days to three weeks. Putting together all of these considerations, we can account for about six orders of magnitude time difference between Kepler and BACON. Since BACON used 1/30 of an hour, and a decade is about 88,000 hours, the ratio of times is about 2.6 million to one, not far from what we would predict.

Similarly, in the case of Planck, we know from documentation that he arrived at the law of blackbody radiation within a few hours of the time when he learned that the previously accepted law (Wien's Law) was at sharp odds with new data in the infra-red range. From some informal and unpublished experiments, we know that some talented contemporary applied mathematicians can solve the same problem (disguised to be unrecognizable) in a few minutes.

These examples do not wholly resolve the problem of speed, which deserves closer examination than it has had. However, it may be less of a problem than appears at first blush; and with Professor De Mey's comments on the cumulative nature of scientific discovery I am in complete accord.

When is it hard to escape from old ideas in the face of revolutionary novelty? Priestly was a holdout against the oxygen theory of combustion, but in his time the phlogisten theory was not as patently wrong as hindsight now makes it. Rutherford's 'discovery' of the electron quickly converted many of the prominent die-hard opponents of atomic theory (e.g. Ostwald), even though it failed to convert others (e.g. Mach). The history of continental drift and plate tectonics shows that views can change very rapidly when the evidence is strong, and can persist for decades when it is equivocal. Scientists may not be as unresponsive to hard facts as the folklore suggests.

I am not enamoured by the comparison of our theory of scientific discovery with cold fusion. Our results are easily replicated; cold fusion has not been. True, "we should be careful not to throw away our belief in the special character of great discoveries too readily". However, among the discoveries we have already explained with our models (Kepler's Third Law, Black's temperature equilibrium law, the discovery of the concepts of inertial mass, specific heat, atomic weight and molecular weight, and many others) are a substantial number that have always been thought "great". Perhaps the reluctance to abandon the "special character" hypothesis is a contemporary example of resistance to revolutionary science, but to suggest that would be to argue ad hominem.

Professor Donald A. Gillies

Most of Professor Gillies objections have already been handled in my general comments. In discussing Kepler, he begins by committing the Fallacy of the Definite Article ("The really difficult part of Kepler's discovery was . . .").

I have already observed that it was no great problem to select period and distance as the two variables. There were few other candidates, and these two had already been proposed by Aristotle. The discussion of the elliptical orbit of Mars is irrelevant to the Third Law, which deals only with average distances. Of course, Kepler started with his acceptance of Copernicus. Science is incremental, and BACON was only seeking to explain the next step, the one Kepler took.

Now we come to Professor Gillies' worry that it would be injurious if my claim were false, but yet widely accepted. Of course, I think the claim is true, but let me for a moment accept his premise. Nothing in our theory would make us reluctant to get help from experts in building an expert system. KEKADA, for example, requires as input information about the expert (disciplinary) knowledge that the scientist whom it is simulating possesses. In

the case of Krebs, we got it from the historian Holmes, but Holmes got it from Krebs' laboratory notebooks and from interviews with Krebs before he died.

Professor Gillies also errs in supposing that we rely upon "the computer's formidable computational powers". On the contrary, a simulation would be unsatisfactory if it demanded more computation than a human scientist is capable of. Our research is aimed at understanding human thinking, not at producing powerful AI systems.

Finally, Professor Gillies simply finesses the whole argument when he warns us that we "will fail to realize that this human creativity is a wonderful resource and one which can be consciously used to produce improvements in the systems of artificial intelligence". First, as just pointed out, the latter is not our goal; and secondly, it is precisely the aim of our research to find out how wonderful a resource, and what kind of one, human creativity is. We wish to study human behaviour without prejudgement of whether its 'wonder' makes it ineffable. That judgement should come at the end, not the beginning, of the research.

Mary Hesse

Professor Hesse starts by asking whether the human mind is *sui generis*. This, she points out, is an empirical question, and the purpose of our research has been to answer it. As far as we have gone, we have shown that the answer is negative.

Professor Hesse observes that the origins of the heuristics incorporated in our programs also require explanation. I agree, and simply note again that explanation is incremental. Discovering the origins of the heuristics is surely, as Professor Hesse says, also a proper subject of research. We have actually done a bit of this (*Scientific Discovery*, Chapter 5) by exploring how making assumptions of symmetry and conservation (of heat) changes BACON's behaviour in discovering Black's law of temperature equilibrium. For this and other reasons, I agree strongly with Professor Hesse that recent historical work on the social construction of theories need not conflict with the conclusions we draw from our programs. My earlier comments on the 'blackboard' are relevant here. Understandably, I do not share the pessimism with which she concludes her comments.

Philip Johnson-Laird

I can only express agreement with Professor Johnson-Laird's comments—up to the point where he introduces his philosophical worries. Specific observations have difficulty in falsifying general explanatory theories; much less difficulty with specific descriptive theories. (Compare Newton's laws with Kepler's.) Whatever the philosophical difficulties, in practice, disagreement about a descriptive theory seldom lasts many years. When it does, scientists go their ways, agreeing to disagree until the issue can be settled by new evidence. There is no compelling reason in science why the fate of a theory must be settled, once and for all, at some given moment.

Consider Prout's law, that every atom is formed of hydrogen atoms. Fractional atomic weights quickly 'falsified' that claim, but on the other side, there were more nearly-integral atomic weights than could be explained by chance. The question remained in limbo until isotopes were discovered, proving Prout 'right'. Then, since agreement with data was still only approximate, atomic packing fractions had to be invented and explained by special relativity and quantum theory. The facts were never much in question, nor the fit of the descriptive hypotheses to the facts. What was in question was what to make of it all.

Simulation of Prout's 'discovery' is not difficult. BACON finds Prout's law if the criterion for noise is sufficiently loose. BACON can also reject the law if the criterion is stricter. To choose, we would need exogenous variables to control the criterion. That would require a more comprehensive theory, and as Professor Johnson-Laird observes, there is still a strategic gap here—room for more doctoral theses to close it.

I agree also with the comment on social psychology, and refer again to the earlier 'blackboard' discussion.

Professor Johnson-Laird also expresses two 'empirical worries'. First, perhaps the simulations of discovery are too coarse-grained, being at the level of strategies rather than the mental processes underlying strategies. The evidence suggests that the adoption and execution of strategies are the principal mental processes involved in problem solving. Of course, there are perceptual and sensory processes also, but we certainly understand a great deal about these from experiment and independent simulations, and they appear not to interact strongly with the strategies in ways that the simulations fail to take into account. I cannot here defend this view at length, but it is defended, implicitly, in the final chapter of *Human Problem Solving* (Newell & Simon, 1972).

The use of production systems and their validity as psychological constructs is based on a great deal of empirical evidence. By contrast, there is, to date, little evidence that the currently fashionable 'connectionist' architectures can account for knowledge that is hard to put in words. Again, a full discussion of the prospective roles of symbolic and connectionist architectures in accounting for cognition cannot be attempted in these remarks.

I should not like to claim that recognition is the only process that is ever labelled 'insight'. Professor Johnson-Laird's examples can be handled comfortably by a combination of a recognition mechanism with problem-solving mechanisms for changing representations and I have already discussed both of these.

I wholly agree with Professor Johnson that imagery is involved in a great deal of problem solving, and would point to the use of imagery in programs like UNDERSTAND and ISAAC, both mentioned earlier, to say nothing of DALTON, STAHL and GLAUBER, all of which can be interpreted as representing chemical structures as mental models. So perhaps we are not in disagreement on these points.

John Losee

Professor Losee is certainly right, that a great deal more work must be done before we know with certainty that all scientific discovery is problem solving. In the meantime, there is not yet evidence that any particular kind of discovery process is not problem solving, and much evidence that many kinds are. It does seem a good strategy to push the strong hypothesis until it is proved false: (Of course, I don't think it will be.)

Some of Professor Losee's discussion and the use he makes of examples, suffer from the Fallacy of the Definite Article. He thinks that whatever DALTON explains leaves unexplained the real nub of theory—in this case, the atomic theory. Of course, that goes back to Democritus, and Dalton did not have to invent it. However, the general answer to his argument is to point again to the nature of explanation as incremental—difference equations can only explain each step as a function of the initial conditions, which sum up all previous steps.

Nor do our models neglect explanatory arguments. DALTON, STAHL and GLAUBER are full of them if BACON is not. Even BACON frequently simplifies the regularities it finds by inventing and introducing into its laws new theoretical (explanatory)

terms such as inertial mass, specific heat or index of refraction. With respect to 'scientific revolutions'. I think I have already had enough to say in my discussion of representation change.

Professor Marconi

What Professor Marconi says about levels of explanation and levels of evidence for explanations seems to me quite correct. It was impossible for me, in my short article, to discuss all the kinds of evidence that have been used to test our models of discovery, both coarse-grained and fine-grained. It is encouraging for the theory that the evidence at all levels, however incomplete, points in the same direction.

However, Professor Marconi occasionally is trapped by the Fallacy of the Definite Article, for example, when he says that "the hardest of such problems is not the spelling out of the rules of analogical reasoning proper ... but the problem of identifying fruitful sources of analogical arguments". The problem of identifying sources of analogical arguments is surely one problem that should be addressed, but it is easy to exaggerate both the role that analogy plays in discovery and the number of analogies that must be searched to find an appropriate one for the given task.

Because research on the modelling of analogical processes is still in a relatively early stage, I should not try to settle the issue here. I will only remark that in analogizing from planetary system to atom, we are using only very general, abstract properties of the former; and at a similar level of abstraction, a pomegranite simply will not do—it will not match the phenomena. As we have a considerable armatorium of methods for tackling problems of recognition and partial matching, I am not dismayed at the task.

Finally, the Fallacy of the Definite Article reappears with Professor Marconi's comment on 'problem raising'. I have already treated both the fallacy and the topic earlier.

W. H. Newton-Smith

It has often been remarked that the same jug of good wine can appear either half-full or half empty, depending on the viewpoint (and the thirst?) of the viewer. The jug that I find half full (or perhaps more than half full), Professor Newton-Smith finds half empty. What I see as replicating some of the most important scientific discoveries in history, he sees as unsurprising success in curve-fitting "for limited and relatively simple cases". Special relativity hasn't been simulated, nor plate techtonics—only Kepler's Third Law, the discovery of the concepts of inertial mass and atomic weight, Black's Law of Temperature Equilibrium, Conservation of Momentum, etcetera.

Nor is there a wine-tasting program, even though there is a program (Cohen's *Aaron*, mentioned above) that produces excellent and aesthetically exciting drawings. True, "what we have been sampling may not be characteristic of the entire enterprise". If the history books are correct in their assessments, it is characteristic of some of the most highly regarded parts of the enterprise.

Next Professor Newton-Smith thinks I commit the Fallacy of the Definite Article. "Scientific discovery is not always a matter of producing a formula which fits the data. It can consist in the recognition of the significance of the formula for explanation and understanding". Indeed, it can and it is one of the tasks of a theory of discovery to explain how explanations and understanding are attained.

In the book *Scientific Discovery*, my colleagues and I describe several computer programs that discover explanations of phenomena. The program STAHL's explanation

and understanding of combustion, first in terms of the phlogisten theory, then of the oxygen theory, are described in Chapter 7. Chapter 8 shows how the program DALTON constructs atomic models for Dalton's data on chemical reactions and genetic models of Mendel's sweetpea data. The KEKADA program not only discovers the role of ornithine in the *in vivo* synthesis of urea, but elucidates the reaction path along which the reaction proceeds—that, in chemists' language, explains the reaction. I cite only examples from the work of our own group, but even these examples show that we can account for many of the processes that are usually called 'explanation'.

In addition, BACON itself provides explanatory content for some of the laws it rediscovers, by introducing (without prodding from the programmer) such theoretical terms as inertial mass, specific heat and index of refraction. It would be interesting for Professor Newton-Smith to suggest what categories of explanation are ignored by these programs so that we can begin to extend our efforts to encompass them.

Finally, Professor Newton-Smith argues that, even if our programs could rediscover all of the discoveries of science, we would still not have a "significant explanation of the process of discovery . . . To achieve understanding we need to know about the mechanism whereby the device converts from input to output". Of course! That is why we are not satisfied to test our programs simply by seeing if they can rediscover historically important scientific laws.

In addition, we have gathered extensive evidence to match the behaviour of the programs, step by step, against the behaviours of scientists they are simulating, sometimes down to the level of laboratory notebooks (Kulkarni and Simon, 1988). Not satisfied with that evidence alone, we have run human subjects in the laboratory, provided with the same data that the scientists had, and have observed their thinking-aloud protocols while they were solving or attempting to solve the same problems. We then compared these laboratory protocols both with the historical records and with the details of the behaviour of our programs (Qin and Simon, 1988).

Data of these kinds show that the simulation programs—our theories—match closely the sequences of human behaviours, thus demonstrating that the mechanisms producing the final outcomes are closely similar in the two cases. In this way we are able to account not only for successes of human discovery, but also for some of the important sources of failure.

Finally, Professor Newton-Smith suggests that only a neuro-physiological theory could provide an explanation of discovery. Of course, since we are not vitalists, we believe that scientific discovery is ultimately the work of the neurons in the human brain. However, explanation does not always or even usually proceed in a single step. I know of no body of theory that explains any biological phenomenon in terms of quarks or other elementary particles. Does this mean that biologists do not believe in the reduction of biological processes to processes among elementary particles? Certainly not. It simply means that reduction must take place in several successive steps, each of which can proceed semi-independently of the others.

I have discussed this question at length in my well-known essay on "The Architecture of Complexity", reprinted in my *The Sciences of the Artificial*, and need not repeat the argument at length here. I will simply make two brief remarks.

First, computer simulation of discovery has demonstrated the sufficiency of certain systems of mechanisms in accounting for discoveries and for the sequence of processes involved in those discoveries. It shows once and for all how mechanism can behave like (creative) mind.

Secondly, we all look forward to the happy day when we will be able to explain how symbolic processes (which we now know so well to execute in computers) can be executed

by systems of neurons. That will be an important achievement in explanation. However, meanwhile, we can continue our progress in explaining how mental processes can be executed by physical symbol systems in ways functionally equivalent to their execution in the brain. It is this kind of explanation that we claim for our theories, and this kind of explanation that has been provided by the information processing revolution in modern cognitive psychology.

I am grateful for the good wishes that Professor Newton-Smith offers to our research venture. Although I am more sanguine than he about how far we have already proceeded along the road, encouragement to continue vigorously is always welcome.

Angelo M. Petroni

Professor Petroni is right in thinking me optimistic. The costs of optimism are low and of pessimism high. Optimism encourages continuing the search, which is bound to lead to something interesting, whether it is what you expect or not. Pessimism leads to inaction and boredom.

Most of the first portion of Professor Petroni's comments have been answered in my earlier discussion of induction. Of course, the programs do induction and their heuristics are inductive, but they do not aspire to inductive proof or certainty, which they surely do not attain. Professor Petroni misses this fundamental distinction between using induction and believing it to be infallible.

The last portion of Professor Petroni's comments expresses preferences about how the word 'logic' should be used. I have stated earlier how I use it, and have shown that my usage agrees with at least one of the dictionary definitions.

Professor Petroni is quite right that the work on discovery leads towards a naturalized epistemology, but wrong in thinking that this excludes a prescriptive point of view. There certainly is a prescriptive theory of football—how to play it—and every coach tries to teach it. Our universities are full of people teaching and learning prescriptive theories of doctoring, engineering, designing buildings, painting, doing mathematics, managing business firms, reading Greek literature, doing biological research—the list is nearly endless. What illusory activity are they engaged in if there is no room for such prescriptive theories? All of Professor Petroni's skepticism would apply, word for word, as readily to these activities as to scientific discovery.

Roger Schank & Lucian Hughes

The comments of Professors Schank and Hughes address one central point—problem finding—and their criticism suffers from the Fallacy of the Definite Article. There is no need to add to what I said earlier about this fallacy and about problem finding.

Guiseppe Trautteur

Professor Trautteur's problems are with the whole Physical Symbol System hypothesis as an explanation of human thinking. His concerns focus on the notions of 'association' and 'denotation'. I will organize my reply around the same notions.

Professor Trautteur challenges us to reduce 'association', 'salience', 'recognition', and so on "to well individuated, conceptually stable constructions based on unambiguously material entities". Exactly such a reduction is provided by list processing languages

such as IPL-V, LISP or OPS-5, the languages in which most of the simulation programs have been written.

The device of description lists (property lists) implemented in all these languages permits them to represent relational structures of arbitrary complexity, hence associations. These languages run on standard computers, in which the symbols and their associations are represented by various kinds of electromagnetic patterns (different in different generations of computers), both computers and patterns being unambiguously material entities.

Salience is a matter of attention. Simulation programs attend selectively to portions of the stimuli available to them, salience being influenced by stored knowledge about objects and by the context defined by goal symbols. As explained earlier, direction of attention is central to the solution of the Mutilated Checkerboard problem. How recognition is accomplished in computer simulations has already been sketched, and is discussed more fully in Feigenbaum & Simon (1984). So I do think such reduction has already been performed and I do rebuke Professor Trautteur in exactly the way he anticipates in his comments.

That brings us to denotation and meaning. Of course, if we agree to incoporate a human agent directly in our definition of meaning, as Derrida in his deconstructionist musings does, the debate is over. However, just why is this legitimate? What can humans do with meanings (or understanding) that computers can't do?

Professor Trautteur complains that Newell, in discussing the Physical Symbol System hypothesis, refers only to 'internal ostension'. In reply, I will describe some computer systems whose symbols have external referents. These systems exhibit the same relation as humans do between their internal thoughts, and their sensory detection and interpretation of an external environment as humans do.

My office overlooks a park with a road running through it. Frequently, I see on the road an autonomous (driverless) vehicle, a van designed and built by our university's Robotics Institute. The van is equipped with cameras that can capture visual information from its environment and with a computer that uses this information to steer the van along the road, avoiding obstacles that may appear. It can usually do this even when there are irregular shadows on the road, fallen leaves or snow. Sometimes it misinterprets the information it senses, gets into trouble and has to be rescued. At present it moves at a good walking pace, but speeds up a little each few months.

The van's computer stores symbol structures that denote various features of the road and the objects on and around it. Information about these objects, originating in photons reflecting from them, is transduced by the van's sensory devices into electromagnetic signals, which then become the symbol structures internal to the computer. This is what the denoting relation amounts to in operational terms—and this is exactly what it amounts to also when photons are transmitted to the human eye from the same scene and transduced there into the symbol structures that inhabit the human brain.

BACON (like the human Kepler) begins with data from the external world that has already been transduced into numerical form. However, the denotation of these data, traced back through the instruments that produced them, is not essentially different from the denotation of the data gathered by the van. No interpreter outside of the system is needed to make them meaningful. The transduction path connecting the internal symbols with the external objects is the meaning.

Understanding consists in constructing an internal representation that retains essential (for whatever purpose) information about the characteristics of the external scene. I have already explained how such understanding is achieved in programs like

UNDERSTAND and ISAAC, and perhaps most relevant of all, Laurent Siklóssy's ZBIE program for using semantic information to learn natural languages.

Professor Trautteur goes on to speak about a "distaste for conditionals" in symbolic processing, which I do not recognize. Turing's 'immediate recognizability', in our present context, is simply the irreducibility of the simplest sensitivities of individual retinal neurons, or their equivalents in a camera. Professor Trautteur is quite right in connecting his qualms with the concerns of Gestalt psychology, phenomenologist philosophy and connectionism (he could have added Gibsonian theories of perception and the new fad of 'situated action'). Careful attention to my last few paragraphs could dispel these concerns.

Physical symbol systems can contain both symbol structures that have only 'internal ostension' and others that have external referents. Hence, there is no difficulty in accommodating both syllables (or phonemes), which belong largely to the former category and words (or phrases, or sentences) which also have external referents. Consequently, physical symbol systems are fully compatible with dictionaries, faithful translations, and the dichotomy between intension and extension, connotation and denotation.

In his penultimate paragraph, Professor Trautteur raises some vague hopes for "the beginnings of an explanation of understanding". We do not need such vague hopes, nor concerns with the limitations of Gödel theorems, which apply to humans as well as to computers. I have just indicated the lines along which an explanation of understanding is already available to us. There is no need to pine for "that still elusive transition between the purely material and the autonomously mental". The Robotics Institute van, not to mention the other programs I have discussed, has already made this transition.

John Watkins

Kant, as quoted by Professor Watkins, is wrong on two counts. There is a logic of scientific discovery, but contrary to Kant, the logic provides heuristics and not infallible rules. However, there is also a logic of artistic creation and it is the same logic. Evidence is to be found in the heuristics of programs that compose music (e.g. The Illiac Suite and the Computer Cantata), and in Harold Cohen's marvelous Aaron program, which produces aesthetically impressive drawings, both representational and non-representational.

However, Professor Watkins' main concern is that our theories of discovery finesse the real problems (that definite article again!) of genuine scientific discovery. First, he points to the role that "essentially novel concepts" may play in important scientific discoveries, and asks whether there could be rules for manufacturing them. Indeed, there could, since BACON has such manufacturing capabilities, and uses them. Given only data about the accelerations of some bodies, it invents and introduces into the statement of laws governing these accelerations the concept (theoretical term) of inertial mass. Given only data on the temperatures of some liquids, it invents specific heat.

Kaplan's program, described in my discussion of the Mutilated Checkerboard, invents a whole new problem representation. A program devised by Weimin Shen (Shen & Simon, forthcoming) invents the concept of genes, and applies it to explaining the inheritance of phenotypic traits, in the manner of Mendel.

All of these examples are reinventions and, hence, do not avoid the danger that Professor Watkins warns against: that essential elements of the answer may have been smuggled into the discovery programs. Let us turn to that question.

First, Professor Watkins suggests that defining the problem may be the real problem. This is by now a familiar argument that I have already refuted. Next, he notes that

BACON had to be provided with "some very broad theoretical concepts". These concepts, of course, are wholly independent of the specific task domains to which BACON is applied; they remain invariant from one domain of application to the next. They are simply the heuristics that allow BACON's search to be selective.

Part of our theory is that scientists can be successful just because they already possess these heuristics when they tackle a new problem and our laboratory experiment on Kepler's Third Law showed that some college students possess them also. The theory does not postulate that an empty head, human or other, can solve problems. Rather, it postulates that human heads are commonly furnished with rather general, but somewhat effective, procedures for solving problems, together with relevant knowledge that gives them expertise in specific domains.

I have already said enough about why BACON was speedier than Kepler to show that illegitimate coaching was not responsible for BACON's success. If "the door is open to various kinds of cheating", the cheating is easily detected by examining the discovery programs, which are open to inspection at any desired level of detail. In *Scientific Discovery* there is enough detail to show just what knowledge BACON has. I think it not implausible that every good scientist has essentially this same knowlege—but this is an empirical question that we have already explored a little, and which is open to further exploration. Cheating can occur only if the critics do not examine the programs diligently. I recommend such activity to them.

Professor Watkins suggests that the scientist cannot know in advance what his goal situation would turn out to be. This is simply wrong. BACON, like Kepler, knew that it was looking for a (relatively simple) mathematical expression that would fit a given body of data well. There are any number of standard tests of statistical fit (BACON employs a very simple one) that will signal when the goal has been attained. They are as available to the human scientist as to BACON.

BACON does not come up with $T^{1.9}=kR^{2.8}$, but if it did, it could use a simple heuristic, that numbers very close to integers should be rounded into integers, to get the 'right' result. This heuristic is used by BACON in other contexts and it is also found in the repertory of every scientist of my aquaintance. There is much historical evidence that it was in Kepler's repertory, too, and it was certainly in Prout's. So if we wish to simulate human scientific discovery, we will surely want to include this heuristic in the program, as one of the scientist's 'initial conditions', and we did. A similar heuristic is also an implicit part of the DALTON program.

Professor Watkins' last paragraph suggests that the real discovery resides in inventing good problem representations. I have already treated this question at length. Hence, I would replace Professor Watkins' future tense with the present tense.

Concluding remarks

I cannot conclude without thanking the commentators for their thoughtful discussions of the logic of scientific discovery. If they have left me largely unconvinced of the errors of my ways, that is not an unusual outcome of a discussion like this.

The final arbiters will be the readers, whose verdict will likely change over the years as new pieces of evidence, on one side or the other, become available. There is no more certainty in this process of discovering a theory of science than there is in the inductions performed by BACON and the other discovery programs. Meanwhile, and whatever that 'final' verdict may be (it is never really final, is it?), the theory that scientific discovery is a

species of human problem solving, not fundamentally different from other species, provides us with an exciting progressive research program. I hope that many will be motivated to engage their efforts in it, wherever it leads.

Acknowledgements

This research was supported by the Personnel and Training Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-86-K-0768; and by the Defense Advanced Research Projects Agency, Department of Defense, ARPA Order 3597, monitored by the Air Force Avionics Laboratory under contract F33615-81-K-1539. Reproduction in whole or in part is permitted for any purpose of the United States Government. Approved for public release; distribution unlimited.

References

KAPLAN, C.A. (1990) In search of insight, *Cognitive Psychology*, 22, pp.374–419.

KULKARNI, D. & SIMON, H.A. (1988) The processes of scientific discovery: the strategy of experimentation, *Cognitive Science*, 12, pp. 139–175.

QIN, Y.L. & SIMON, H.A. (1990) Laboratory replication of scientific discovery processes, *Cognitive Science*, 14, pp. 281–308.

SHEN, W. (1984) Functional transformation in AI discovery systems, *Artificial Intelligence*, 41, pp. 257–272.

SHEN, W. & SIMON, H.A. (1992) Fitness requirements for scientific theories containing recursive theoretical terms, *British Journal for the Philosophy of Science* (forthcoming).

SIMON, H.A. (1981) *The Sciences of the Artificial*, Cambridge, MA, MIT Press (2nd edn).