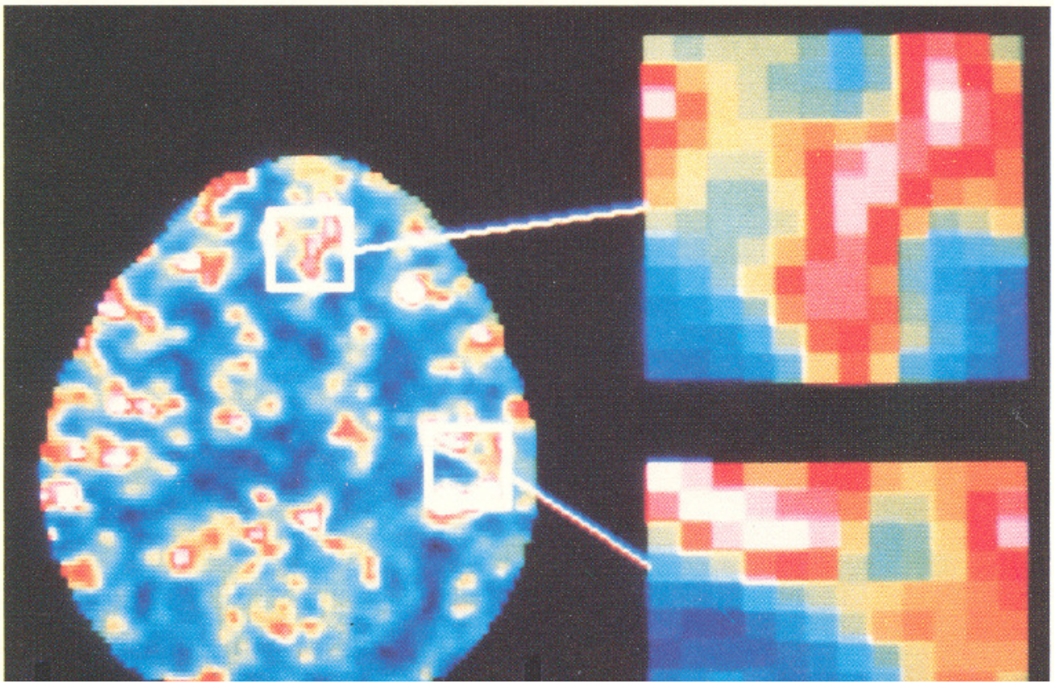
CODEN MMACEO
ISSN 0924-6495

MINDS and MACHINES

Journal for Artificial Intelligence, Philosophy, and Cognitive Science

Vol. 14 No. 1 February 2004



A.I., SCIENTIFIC DISCOVERY AND REALISM*

MARIO ALAI
Istituto di Filosofia, Università di Urbino, via Saffi 9, 20129, Italy; E-mail: mario.alai@libero.it

Abstract. Epistemologists have debated at length whether scientific discovery is a rational and logical process. If it is, according to the Artificial Intelligence hypothesis, it should be possible to write computer programs able to discover laws or theories; and if such programs were written, this would definitely prove the existence of a logic of discovery. Attempts in this direction, however, have been unsuccessful: the programs written by Simon's group, indeed, infer famous laws of physics and chemistry; but having found no new law, they cannot properly be considered discovery machines. The programs written in the «Turing tradition», instead, produced new and useful empirical generalization, but no theoretical discovery, thus failing to prove the logical character of the most significant kind of discoveries. A new cognitivist and connectionist approach by Holland, Holyoak, Nisbett and Thagard, looks more promising. Reflection on their proposals helps to understand the complex character of discovery processes, the abandonment of belief in the logic of discovery by logical positivists, and the necessity of a realist interpretation of scientific research.

1. The problem of Scientific Discovery

Is Scientific discovery a rational activity, is there a method for it, or a "logic" to be followed? There has been a long debate on this question among philosophers of science, with people like Francis Bacon, John Stuart Mill and Hans Reichenbach answering "yes", and no less important characters, such as William Whewell, Albert Einstein, Carl Hempel, and Sir Karl Popper, answering "no".

No wonder it is so, since the question is torn between the horns of a seemingly inescapable dilemma: on the one hand, discovery must be rational, for we honour great discoverers like Newton, Lavoisier, Einstein, etc., as exceptionally rational minds, not as lucky, or sensitive, or super-naturally endowed people. But if the process of discovery is rational, mustn't it therefore follow rational criteria and rules, hence a logic? On the other hand, it is well known that chance, luck, and insight often play an important role in discovery (as it happened, for instance, with Kekulé's discovery of the hexagonal structure of benzene, which was prompted by a dream).[1] And above all, if discovery were just a matter of rule following, why couldn't anyone learn the necessary rules and become a great scientist? Or why couldn't the scientists themselves just follow the logic of discovery and program in advance new discoveries, and rapidly achieve such results as a cure for cancer, or the cold fusion of atom, which while sorely needed still elude the efforts of researchers?[2]

It might be thought that a logic of discovery is a necessary but not sufficient condition for the success of research; still, no one so far has

Minds and Machines **14:** 21–42, 2004.

satisfactorily shown what this logic is, or codified its rules. No doubt, Bacon and Mill provided clear examples of such rules, respectively through the "tables" of presence, absence and degrees, and through the canons of concordance, difference, concomitant variation and residues, and something similar did Whewell and John Herschel. But these are simply guidances to induction as empirical generalization, allowing to discover empirical regularities or positive correlations among known factors. As acknowledged by Mill,[3] they have little to say when it comes to the discovery of unknown entities or forces, of representations of unobservable levels of reality, or of unifying theories. In other words, they have little to say concerning the most important discoveries, such as the theory of gravitation, the atomic theory, the electromagnetic theory, the relativity theory, and so on. The same is true, *a fortiori*, of Reichenbach's method for searching for a limit of relative frequencies.[4]

Peirce's abductive logic, on the other hand, supposedly leads to discover causes or explanations beyond the level of empirical data;[5] but as pointed out by Laudan and Pera,[6] abduction does not lead to the idea of the relevant concepts or hypotheses, but rather presupposes it; abduction, therefore, is to be considered more a logic of *pursuit* than of discovery.[7]

If one accepts the hypothesis that machines can be built to perform the intelligent tasks of the human mind,[8] one *ipso facto* sides with the supporters of the logic of discovery: for on the one hand discovery is certainly one of the tasks of human intelligence, and on the other hand mechanical processes are algorithms, and algorithms follow a logic. Thus, if a machine could run a process of discovery, there would be at least one logic of discovery, the one followed by that machine. Supporters of the hypothesis that machines can be built to perform the same tasks *by the same processes* as human intelligence[9] would claim even more, namely that the machine's logic of discovery could also be the same logic followed by human researchers. Moreover, Artificial Intelligence would seem here to offer a unique opportunity for empirical testing of a philosophical conception: believers in the logic of discovery can prove their point simply by producing a machine that actually makes discoveries. Needless to say, people have seized this opportunity, even if judgement on their outcomes crucially depends, as it turns out, on what is meant by "actually making discoveries".[10]

2. The Bacon Programs and Simon's Approach

This is precisely what Herbert Simon (philosopher of science, computer scientist and Nobel laureate for Economics) tried to do, together with his co-workers (among whom Pat Langley, Gary Bradshaw and Jan Zytkow authored with him the panoramic volume *Scientific Discovery. Computational Explorations of Creative Processes*).[11] They produced computer programs intended to show that scientific discovery is just *problem solving*, that problem solving is a rational

Minds and Machines **14:** 21–42, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

activity, that it has a logic, and that it can be mechanized. In fact, they took this challenge so seriously that they chose some of the most famous discoveries in the history of science, in order to prove that a machine could do the same.

The first series of their programs is significantly called "Bacon" (with versions from Bacon$_1$ through Bacon$_5$), after the philosopher who thought of mechanizing discovery to the point of making it almost independent of human skill.[12]

Bacon$_1$ looks for laws describing regularities in its body of data, following rules ("heuristics") such as:

- if the value of a term is the same in all data clusters, assume that it is constant;
- if the values of two terms are linearly related in all data clusters, assume that such relation is constant;
- if the values of one term increase as the values of another term decrease, consider their product, and see whether it is constant;
- if the values of two terms increase together, consider their ratio, and see whether it is constant.[13]

In this way, Bacon$_1$ is able to discover Kepler's third law starting from the values of periods and distances of planets from the sun; Boyle's law starting from the values of pressure and volume in a gas; Galileo's law of uniform acceleration starting from times and distances; Ohm's law starting from the length of a wire and the intensity of current.

Bacon$_2$, instead, searches for constant derivatives, rather than products and constants. Bacon$_3$ applies the same heuristics as Bacon$_1$ to bodies of data including more than two independent variables, thus discovering the ideal gas law from the values of temperature, pressure and volume for different quantities of a gas; Coulomb's law from the values of two charges, distance and force acting between them; and more complex versions of Kepler's third law and Ohm's law.[14]

Bacon$_4$ can deal with cases in which data are names of objects, rather than numerical values for their properties. Thus, it can rediscover Ohm's law when provided just with the names of different wires and batteries, and the values for intensity of current. It does so by adding to the heuristics of Bacon$_1$ and Bacon$_3$ a new one, to the effect that when the values of a given property of one object (e.g. the current of one battery) vary with the related objects (e.g. with different wires), then it must be postulated that there is a property of these related objects which is responsible for such variation (e.g., conductance), and whose values are proportional to those of the first property (current).[15]

In this way, Bacon$_4$ postulates the existence of new unobservable properties, whose effects are supposedly manifested by the available data. This happens also when it finds out about volume, density, index of refraction, specific heat, gravitational mass and inertial mass. The authors stress that this amounts to the introduction of *theoretical*, i.e. unobservable, properties, whereas an observable property would be one that may be observed or measured either without instruments, or by instruments which are not considered problematic in the given contest. In the case of Ohm's law, for instance, data are the names of objects and the values of intensity of current, hence an amperometer might be considered non problematic; but conductance cannot be measured either without instruments or with the help of an amperometer.[16]

Nonetheless, this does not seem to be a full-fledged example of theorization, since the new property is not embedded in a whole theory describing it, its nature, causes, functioning, effects, etc. Its introduction, here, is just matter of detecting a regularity in the behaviour of observable objects and properties, and ascribing to it as a cause a property which is identified just by means of this particular effect (exactly as one could say that since opium makes people sleep, it must have a *virtus dormitiva*)[17]. In fact, its values are computed directly from the values of the observable properties involved.[18] Thus, we should say at most that $Bacon_4$ takes some of the most elementary steps in the process of theorizing.

Another heuristic tells $Bacon_4$ to look for common divisors and their regularities. In this way the program may be applied to chemical research, and starting from the proportions of weights and volumes of elements in compounds it finds the molecular and atomic weights of various elements and compounds.[19] This is not to say it discovers either the molecular or atomic theory, however; for it finds just numbers, and it is only in the light of our knowledge of atomistic chemistry that we may interpret those numbers as atomic or molecular weights.

The last version, $Bacon_5$, was created by adding to the earlier ones the notion of physical symmetry, and the rule that if a particular relation holds among a set of variables, (e.g., two objects with respective initial and final velocities), it must be presumed that it holds among variables of the *same type* (e.g., two different objects with respective initial and final velocities). Hence, on condition of being told which variables are of the same type, the system is able to speed up significantly its search for regularities (such as the law of conservation of momentum, Snell's law of refraction, or Joule's law). $Bacon_5$ is thus *theory-driven*, i.e. it imitates those cases of actual science in which research is not merely based on data, but on theoretic presuppositions as well.[20] Even more theory-driven is a different system, Black, which has inbuilt the notion that certain properties are conserved, by which it can find the law of specific heat much faster than $Bacon_4$.[21]

In spite of their vast potentialities, the Bacon programs cannot deal with qualitative laws, nor give structural descriptions of reality. This greatly limits their applicability, especially to the field of chemistry, and to overcome this weakness three new programs have been developed: Glauber, Stahl and Dalton.[22] Glauber uses as data descriptions of chemical reactions (such as "hydrogen chloride reacts with ammonia to form ammonium chloride") and properties of chemicals (such as "hydrogen chloride tastes sourly"), and its heuristics instruct it to group in the same class the chemicals entering the same type of reaction, or having the same properties. Thus, it forms the classes of salts (tasting salty and formed by the reaction of an acid with an alkali), of acids (tasting sourly and reacting with alkalis to give salts) and of alkalis (tasting bitter and reacting with acids to form salts).

Stahl uses the same kind of data about chemical reactions to detect the components of various substances: if **x** and **y** react to produce **z**, it infers that **z** is composed of **x** and **y**; the same if **x**, **y** and **w** react to produce **z** and **w** (in this case it infers that **w** is idle in such a reaction). Furthermore, it draws inferences by substituting substances with their components, and *vice versa*. For instance, if **z** is composed of **x** and y, and **z** reacts with **q** to give **r** and **y**, Stahl infers that **x**, **y**, and **q** react to produce **r** and **y**; hence (discarding **y** that is idle), **r** is composed of **x** and **q**.

In this way, if fed with data about reactions interpreted as the phlogiston theorists did (e.g., charcoal and air react to give phlogiston, ash and air), it infers the same analysis of substances given by the phlogiston theory (e.g., that charcoal is composed of phlogiston and ash). In fact, it stumbles on some of the same problems that puzzled the phlogiston theorists, and that lead to modifications in their theory: when fed with the results of reactions conducted by Priestley in 1773, it is forced to conclude that one of the components of mercury is mercury itself! On the other hand, when fed with reactions as interpreted in the oxygen theory, Stahl correctly yields the nowadays accepted composition of chemical substances.[23] Nonetheless, it is clear that Stahl does not discover either the phlogiston or the oxygen theory, but simply *applies* them: it accomplishes what Kuhn would have called normal science tasks,[24] or perhaps a simple analytical elaboration of data.

The same can be said about Dalton: starting again from data about reactions, and assuming that the number of molecules involved in a reaction is proportional to the volumes of the respective elements, it infers the molecular structure of compounds. Moreover, assuming that atoms are conserved in reactions, and that molecules are composed of the smallest number of atoms compatible with the law of conservation and the known molecular structures, Dalton infers the atomic structures of elements and compounds.

3. **The Turing Tradition**

Among the many critics of Simon's work[25], Donald Gillies points out[26] that his programs can simulate the discovery of *known* laws, but have not been able to discover any new law, or solve any practical problem. But obviously, finding what is already known is not making a discovery at all! (Moreover, as I have stressed, they haven't found any *theory*, not even old ones). Gillies contrasts them with the programs produced by disciples of Alan Turing and researchers working in their tradition, such as Ehud Shapiro, Stephen Muggleton, Donald Michie, Edward Feigenbaum, Bruce Buchanan, J.R. Quinland and Ivan Bratko, programs which have successfully been applied to the solution of practical problems and have made *new* (hence *actual*) discoveries.[27]

For instance, DENDRAL, developed since 1965, accomplishes what an expert chemist might do, inferring the molecular structure of organic compounds from their mass spectrogram. ASSISTANT has been able to diagnose various kinds of disease better than human specialists. GOLEM was able in 1991 to predict the secondary structure of proteins from their primary structure. Primary structures, in fact, are easily known, but secondary structures are more important, and up to then they could be discovered only by long and expensive methods, such as nuclear magnetic resonance or X-ray crystallography. GOLEM, instead, discovered a number of rules linking certain primary structure characters to certain secondary structure characters, such as:

> There is an α-helix residue at position B in protein A if:
> 1) the residue in B-2 is not proline,
> 2) the residue in B-1 is neither aromatic nor proline,
> 3) the residue in B is large, non-aromatic and non-lysine,
> 4) the residue in B+1 is hydrophobic and non-lysine,
> 5) the residue in B+2 is neither aromatic nor proline,
> 6) the residue in B+3 is neither aromatic nor proline, and it is either small or polar,
> 7) the residue in B+4 is hydrophobic and non-lysine.[28]

Gillies credits such rules as “new laws of nature” discovered by GOLEM. But it is far from clear that they qualify as such, because (a) they have a very low level of generality, as the above example shows, (b) as he himself concedes they fail in about 20% of cases, and therefore (c) it is dubious that they describe actual causal relationships, as opposed to contingent statistical correlations.

Summing up, all these programs produce generalizations connecting a target property (such as a secondary structure character, or a particular disease), to the presence or absence of symptomatic properties (such as primary structure characters, or symptoms manifested by patients); they do so by checking which symptomatic properties are present or absent when the target property is known to be present. This is to say, they practice, though with sophisticatedly iterated procedures, nothing but Baconian or Millian induction. No wonder therefore these programs have only discovered low level generalizations, and no theories, theoretical laws, entities or explanations. Just as enumerative induction, also Bacon’s tables and Mill’s canons, in fact, may establish horizontal links among empirically known entities or properties, but no vertical links among observable and non-observable levels of reality.

According to Gillies the difference between the approach of Simon’s group and that of Turing’s tradition, allowing the latter to achieve new discoveries (aside from the fact that the former tries to repeat historical discoveries, while the latter tries to solve open problems) is double: first, Simon’s approach is “psychological”, i.e. it imitates human inferences, while that of Turing’s tradition is “logical”, i.e. it starts with logical (inductive) inference rules and sees what they can lead to; second, they differ just in the way intuitive pre-Fregean inferences differed from formalized Fregean inferences.[29]

Actually, however, neither of these differences seems to hold: Simon's inference procedures are obviously formalized, since they can be carried out by a machine; and the procedures of Baconian induction implemented in the Turing tradition programs are no less instances of actual human inferences than Simon's heuristics. The only real difference seems to be that Baconian induction is a very general (although weak) inferential mechanism, while Simon's heuristics are stronger, but therefore also of narrower applicability. Hence, it is hard to apply them to problems for which a clear solution strategy is not already known. Simon could certainly have written programs to achieve results of the kind of DENDRAL, ASSISTANT or GOLEM. Simply, he wouldn't have thought that such results could teach much on the problem of the rationality and mechanizability of scientific discovery.

4. **An Alternative Research Program**

Up to this point, at any rate, the attempts based on Artificial Intelligence might seem to yield a negative answer to the question of the logic of discovery, quite against the hopes of their authors! For it would seem that scientific discovery (at least in its most significant instance, theoretical discovery) cannot be mechanized; hence, that it does not have a logic, hence, that it is not rational. Actually, these would be fallacious inferences, because there might be a logic even without explicit rules and mechanizable algorithms; and there can be rationality (prudential or argumentative rationality, as advocated since Aristotle and up to Thomas Kuhn)[30] even without logic. Still, even if not yielding a negative answer, Artificial Intelligence would seem utterly unable to yield a positive one.

This conclusion could be avoided by suggesting that perhaps so far all the attempts of mechanizing discovery have failed because wrong-headed, and new attempts might succeed by adopting different approaches. An example could be the proposal advanced by John Holland, Keith Holyoak, Richard Nisbett and Paul Thagard (henceforth HHNT) in *Induction. Processes of Inference, Learning and Discovery*.[31] They try to reconstruct the cognitive processes by jointly relying on cognitive science, philosophy and computer science, and describe a program ("PI", for "processes of induction") which purports to replicate the crucial features of such processes.

This approach has not yet developed into a full-fledged research program like that created by Simon and his group, who by a huge investment in terms of time and efforts produced such and articulate series of more and more complex programs; nor HHNT's program has had practical applications like the expert systems in the Turing tradition, yet, due to its greater complexity and newer conception. Hence, this alternative approach is still at a largely programmatic stage, and the program PI has not yet made any new discovery, nor as many and as complex old ones as those achieved by Simon's programs; yet, it embodies new stimulating ideas, and, compared to earlier attempts, it appears quite promising for a number of reasons to be examined in the following sections.

Here, it may well happen as for Kuhnian paradigms:[32] at the beginning, a new paradigm exhibits more promises than actual achievements; it certainly doesn't have as many confirming experiences and successful applications as those yielded by its older competitor in years of "normal science". Still, its sketchy outlines and a few exemplar solutions may be enough to give the sense of a new attractive way of looking at things, leading scientists to a significant "Gestalt switch". For those who come to appreciate it, this new perspective in itself may be more important than the fact that it still needs to be fully worked out and applied to problem solving, for they are confident that this task can be accomplished in a more or less routine way. Without such a confidence, it is hard to see how new paradigms could attract enough interest, support and resources to eventually yield their remarkable successes.

Since the publication of this book, the four authors didn't publish any further joint work, but they went on exploring developments and implications of the same general approach for different specialized topics. Such further researches resulted in some more recent books, written individually, in couple, or jointly with different people. For instance, in Thagard (1988) PI is newly discussed with a sample run applied to the wave theory of sound. Thagard (1992) discusses conceptual change in scientific revolutions by the aid of PI and ECHO (Explanatory Coherence by Harmony Organization), another connectionist program; these programs yield here interesting reconstructions of cases such as Lavoisier's chemical revolution, Darwin's revolution, and Wegener's geological revolution. Holyoak and Thagard (1995), devoted to analogy in creative thinking, introduces new programs as ACME (Analogical Constraint Mapping Engine) and ARCS (Analog Retrieval by Constraint Satisfaction). Still further contributions can be found in Holyoak and Barnden eds. (1994), Thagard (1996), Magnani, Nersessian and Thagard eds. (1999), Gentner, Holyoak and Kokinov (2001).

While this literature shows that, because of the complex problems tackled, HHNT's research program is not ready to yield practical applications, yet, it also shows that it has good chances of progressive expansion[33] and fruitful development (in fact, it might be suspected that the modest success of Simon's programs is due to the attempt to apply artificial intelligence to such a difficult question as scientific discovery before it reached sufficient maturity). On the other hand, since here we are more interested in the critical appraisal of the general approach as such, than of its recent developments and specializations, the original work (*Induction. Processes of Inference, Learning and Discovery*), with its synthetic account, may still offer the best vantage point for our discussion. I shall then refer to it in outlining the basic ideas of this approach.

Knowledge is represented by HHNT as the construction of mental models of the environment.[34] Environment, in turn, is described as consisting of states and transition functions among them: for instance, the fact that all bodies moving at time **t** have ceased to move at time **t'** may be described as the state **S(t)**, in which bodies move, the state **S(t')**, in which bodies stand still, and the transition function **T**, turning whatever moves into something still. Of course, there can be long chains of states and transition functions, and particularly important chains are those in which the initial state represents a theoretical or practical problem for the knowing subject, and the final state represents its solution (HHNT stress the importance of pragmatics in their approach).

For HHNT the mind is like a bulletin board, on which "messages" (i.e. propositions) are posted. More precisely, one could speak of two sides of a bulletin board: a front side, on which currently active messages are posted (i.e. propositions in the working memory, currently involved in cognitive processes), and a back side, with non-active messages (propositions stored in the long term memory). In mental models states are represented by categoric messages (e.g.: "all bodies move"; "this body is moving"; etc.), and transition functions by hypothetical messages, or "rules" (e.g., "if a body moves at time **t**, it will stand still at time **t'**").

Beside *empirical* rules, describing transition functions in the environment, there exist also *inferential* rules, governing the construction and modification of models in the face of inputs from the environment. These include generalization rules, simplification rules, specialization rules, etc. A generalization rule, for instance, prescribes that if a message is posted to the effect that all members of a sample of objects of type **A** have property **P**, then another message should be posted to the effect that all objects of type **A** have **P**. A simplification rule says that if we post an empirical rule to the effect that all feathered and large-winged animals fly, and another to the effect that all feathered and small-winged animals fly, we should post another one to the effect that all feathered and winged animals fly. A specialization rule prescribes that if a counterexample is found to an empirical rule, the latter should be replaced by another rule allowing for that exception.

Empirical rules and categoric messages typically form chains: suppose a message is posted as a result of an experience input (e.g., "**x** barks"). If a rule is also posted having the content of such message as its antecedent (e.g., "if **x** barks, then **x** is a dog"), then a message stating the consequent will also be posted ("**x** is a dog"), and so on. Actually, a message may be posted by the converging effect of more than one rule (e.g., "if **x** is a fox-terrier, then **x** is a dog", etc.), and in turn it may start many chains at once, depending on the activated rules (e.g., "**x** barks"; "if **x** barks, then **y** will wake up"; "**y** wakes up"; etc.). Thus, an ever-growing number of connections is activated, in a process of *spreading activation*. In fact, the same mechanism accounts also for a spreading activation of concepts: the concepts involved in an active rule become active too, and each of them in turn activates all the other rules in which it is involved, which in their turn activate other concepts, and so on. (Concepts themselves are conceived by HHNT, more or less as in Putnam's account of scientific concepts, as clusters of rules.[35] A rule might connect two concepts, for instance, by expressing relations of similarity (cats are like tigers), of causations (smoke causes lung cancer) or of category (cats are mammals)).

As it happens, however, there is only limited room on a bulletin board, as well as in the human mind. Therefore, not all potentially activated messages will be posted, and messages will *compete* for room on the board. Besides, mutually incompatible messages that might be posted will be in competition even independently of room scarcity. The competition takes place more or less as it happens on the *economic market*: messages and rules in a chain can be viewed as suppliers, middlemen and consumers, where selling and buying consists in activating or being activated. When the final ring of a chain constitutes the successful solution to a problem (hence the importance of the pragmatic dimension in this approach), its success is comparable to profit, which is duly shared with each preceding ring, as each buyer pays back the goods or services that reached him through the chain.

Such "profits" increase the "capital" of each message in the chain, i.e. its strength in the competition (or, out of metaphor, its credibility). Reversely, when the final ring is unsuccessful (as it does not solve the problem, or it is refuted by experience), it loses part of its capital, and so do all the other rings in the chain. In the competition for posting, the winner is determined by its capital and by the total capital of all the rules and messages concurring to its activation. In practice, the environment's feed-back gradually reinforces successful beliefs and extinguishes unsuccessful ones, just like market reactions make efficient firms flourish and inefficient ones go bankrupt.

When it comes to scientific discovery, it may concern either laws or theories, which in HHNT's account means either rules or models. New rules[36] may be generated either from old rules or from data. New rules may be generated from old rules, for instance, by applying the inferential rules of simplification or specialization: from a rule with unnecessarily complex antecedent (e.g., "if **x** is feathered and large-winged, then **x** can fly") a new one is generated with a simpler antecedent ("if **x** is feathered and winged, **x c**an fly"); from a rule to which there is a counterexample, a new rule is generated including the counterexample as an exception case.

Otherwise, new rules may be generated from the data, either by applying a generalization rule, for instance as in enumerative induction, or by *abduction*. If we observe that various objects with property **P** also have property **Q**, by a generalization rule we may generate the rule "If **x** is **P**, then **x** is **Q**"; instead, if we wish to explain why a given object has property **Q**, we may notice that it has also property **P** (and perhaps that even other objects with **Q** have also **P**); then we may form the same rule, not so much on the basis of the various observed instances, as because such rule, coupled with the fact that the given object has **P**, would explain its having **Q**, thus reasoning by abduction.[37]

All this clearly concerns empirical rules. HHNT do not discuss the generation and evolution of inferential rules; but on the one side it is apparent that inferential rules are much fewer and less changeable then empirical rules; on the other side, it does not seem impossible that in the long run some new inferential rule may be generated, and may gain or lose strength by the same mechanisms as empirical rules, i.e. by their contribution to the final success or failure of a chain of messages; after all, according to HHNT, the distinction between empirical and inferential rules is not a sharp one.[38]

While interpreting scientific laws as rules, HHNT interpret theories as models (that is, whole systems of rules and categoric messages), and suggest that analogy plays the central role in their discovery.[39] This is just natural, since theories typically model non-observable systems, and how else could one figure out about unobservable structures, except by analogy with the observable ones?[40] If finding a theory for a particular set of phenomena is to embed them in a model, working by analogy means taking three steps: first, finding an already accepted model of a different set of phenomena as the convenient *source* for the new model; second, mapping the various aspects of the phenomena we are investigating onto aspects of the phenomena of which we already have a model; third, constructing the new *target* model by positing objects, properties, relations, etc., as corresponding *via* the above mapping to those of the source model.

(Since the source model must already be part of our knowledge, the question may arise, how it was built. If it concerns non-observable phenomena it was likely built by analogy, too. If it concerns observable phenomena, it was built by empirical inputs and inferences. Since a model of observable phenomena may be the source for an analogical model of non-observable phenomena, all models are ultimately based on empirical inputs).

All this is convincingly rendered in HHNT's spreading activation account of cognitive processes: as we noticed, each concept is linked to other concepts by rules stating the existence of similarity, causality, or categorial relations between them; now, supposing we are trying to embed in a theoretical model the phenomenon that **a** and **b** are related by relation **R**, each of the concepts **a**, **b** and **R** will be activated, and therefore cause the activation of the various different concepts linked to them in the above ways. In turn, each of the newly activated concepts will activate others, and so on. We may find then that there is a concept **m** related to **a** just in the way a further concept **n** is related to **b**, and there is a relation **Q** obtaining between **m** and **n** just as **R** obtains between **a** and **b**. If we already have a full model in which **mQn** is embedded, i.e. if we know the full net of causal relationships of **mQn** to other concepts, this may work as a source model for **aRb**; moreover, at this point we already have a mapping between the phenomenon to be theorized (**aRb**) and the source model, and by extending such mapping we may begin the construction of the target model. In terms of the bullettin board metaphor, this is to say that there is a non active model on the back side, that thanks to its connections to currently active concepts (a) becomes active, and (b) qualifies as a possible source model for the phenomenon to be explained.

An example is supplied by the discovery of the ondulatory theory of sound, analyzed in Vitruvius' writings: we wish to develop a theory of sound, explaining among other things the fact that sound is reflected, or that it propagates. The concept of reflection activates the concept of water waves and of rope waves, for they reflect, too. Messages like "rope waves reflect" and "water waves reflect" are then posted. As a consequence, also the generalization "all waves reflect" is posted. But this generalization could explain why sound reflects, if sound was a kind of wave; hence, by abduction, the system goes on posting "sound is a wave". More or less the same process would follow from the activation of the concept of propagation. Otherwise, the concept of sound might activate that of musical instruments, and the latter the concept of string instruments. "String instruments" might then activate "ondulatory vibration", and this again would connect the idea of sound to that of waves. (Of course, many other links would be activated as well, but eventually many of them would prove useless and be disactivated).

PI, HHNT's program, instantiates the spreading activation model of the mind and reproduces simple processes of discovery like the just mentioned one. We shall now examine some of the reasons why it can be considered a promising alternative to the older attempts at mechanizing discovery.

5. Serial vs. Parallel Computing, Complexity, and the Lesson on Scientific Discovery

How does HHNT's strategy differ from those followed by Simon and by Turing's disciples, and why should it fare better in the attempt to mechanize discovery? In order to answer these questions, it should be recalled that up to now computing has mainly been performed by serial, or digital computers, direct instantiations or descendants of the Turing machine. Yet, the alternative option of connectionist machines or neural networks, i.e. parallel or analogical computers, was considered by Turing himself,[41] and is arising more and more interest today. This distinction matters to our problems, in the first place because the brain is a neural network, after all, hence it is a plausible suggestion that neural networks or connectionist machines have the best chances of reproducing cognitive processes. Now, HHNT's approach is basically a connectionist one (though with differences and qualifications).[42]

A serial computer goes through one state after another; with respect to any given state it either is or is not in that state, without further alternatives; it uses a well defined set of data as an input, and given those data its output is easily predictable: hence, in the case of scientific research, it may be used to compute complex numerical or qualitative relations among known factors, but it cannot discover structures or mechanisms genuinely unpredictable at the moment of its programming. This, as we noticed, is the limit both of Simon's programs and of the programs of Turing's tradition.

On the contrary, the mind may be in many states at once (it may entertain different beliefs or propositional attitudes, perceptual states, etc.), and so can connectionist systems: many messages and many chains can be activated at once in HHNT's model. Moreover, the mind is gradual both with respect to a single state (we may have weaker or stronger beliefs, for instance) and to an alternative between states (we may believe **P**, or believe **P** more than **not-P**, or be half-way); the same happens in HHNT's system, thanks to the varying strength of messages, and to the gradual overturning of one message or chain by another. Again, mental processes are not closed elaborations of a restricted body of data, but open, at any time, to the influence of a potentially endless number of inputs, so that the final state is in no way predictable beforehand. Even this feature seems to be captured by connectionist machines, as well as by HHNT's model of the mind. Thus, it is at least conceivable that systems of this kind make genuinely unforeseen discoveries.

Thus, mental processes, and discovery processes in particular, seem to exhibit a higher degree of complexity than traditional computers; connectionist systems, on the other hand, appear to have better chances to achieve a similar degree of complexity. This may then bear on an interesting historical problem: why, after the idea of a logic of discovery had been supported by such epistemologists as F. Bacon and J.S. Mill, was it abandoned between '800 and '900, especially by the logical positivists? And why is it becoming popular once again in our days?[43]

Part of the answer may be that the logical positivists' main problem was to establish a neat demarcation between science and metaphysics, and they solved it by a powerful and very simple criterion, on which their whole philosophy was based: verification. However, that criterion was too simple, for (as it was to become clear later on) it could not yield a full account of the meaning of scientific terms, nor of the justification of hypotheses. In any case, it was obvious even to them from the beginning that discovery could not be captured by such a simple philosophy, and that may explain why they excluded it from their interests and from their tasks: discovery is a holistic procedure, in which the potentially endless aspects of a complex environment continuously interact with an equally open-ended endowment of conceptual structures, background knowledge, methods and criteria.

With the liberalization of logical positivism, and even more with its final abandonment, it was recognized that meaning and justification were complex matters, and then even the equally complex matter of discovery could become again a legitimate question in philosophy of science. Still, a satisfactory way to deal with such complexity has yet to be found. Acknowledging the holistic and complex character of science in general, Feyerabend concluded that there is no logic in science, and in discovery in particular. But even on more moderate views discovery (as opposed to justification) is too complex to be governed by logic or rules.[44]

On the contrary, both Simon and the researchers in Turing's tradition have gone back to the faith that science (discovery, in this case) can be analyzed into simple processes (Simon came from the logical positivist tradition, after all). But once again, the simplicity approach has failed: in Turing's tradition only the elementary methods of Baconian induction are used, which allow just a very low level of discovery. Even Simon has relied on relatively simple heuristic methods, but devising a different specific mechanism for each discovery, and obviously taking inspiration from his knowledge of the procedure historically followed or of its results. Thus, his programs have been able to make significant "discoveries", but only old ones. Following his approach, one could try to make *new* discoveries only by writing a program that included a specific heuristic for each possible discovery, i.e. potentially infinite heuristics, and running all of them each time.

The only possible alternative, if there is one, is a program that can devise by itself the discovery path required by the solution of each specific problem; that is to say, a really *intelligent* program. If for instance we could program a robot to turn the tap, collect water in a tub, pour soap, dip the clothes, etc., by linking such procedures in a chain we could get it to do our laundry; yet, we wouldn't say it knows how to launder. We would grant that it knows how to launder if just upon being ordered to launder it could go on by himself, choosing the procedures and adapting them to the situation (there may be no taps, for example, and then water may have to be drawn from a well, etc.). In this case the robot would be more intelligent than in the first case; and it would be even more intelligent if we could just ask it to look after the house, and let it find out by itself whether to launder, or iron, or clean the floor, etc., and how. This is to say, intelligence seems to involve and to be proportional to the ability to pursue ultimate goals by flexibly and adaptively choosing the means or the intermediate goals.

Now, it seems that this is precisely what a machine needs in order to do genuine research, without being previously instructed on how to make each discovery: it must be able to pursue the goal "discovery". Further, it seems that this is precisely what HHNT try to do, where they basically differ from both Simon and Turing's tradition, and why their attempt is likely to be more successful: they try to devise a mechanism intelligent enough to pursue the goal of discovery. It is obvious that only such a mechanism may qualify as a plausible imitation or reconstruction of human abilities, and motivate the claim that scientific discovery is a logic or rational process in a significant sense.

This is why I think such a connectionist approach looks promising. It might be objected that the connectionist character cannot make the real difference, for anyway a connectionist machine can be instantiated by a Turing machine. But first, I am not suggesting that the difference is made by the connectionist character alone, but together with the adaptive character; second, in any case, what matters is the structure of the resulting procedure (i.e., that it be parallel, open to endless inputs at all times, complex and self-correcting, etc.), rather than the underlying mechanism implementing it: if we can get a Turing machine to work that way, so much the better!

6. Discovery and Realism: the Role of Models

To better appreciate the reasons of HHNT's superiority, we may ask what makes it possible, i.e., what enables PI's search for discovery to be self-directed. The answer involves at least the following two features: first, HHNT give a correct and fruitful characterization of what discovery consists in and how it is achieved, namely, the construction of *models* of reality; second, they let nature itself steer the program toward its goal, through a feedback mechanism. Since discovery is by definition finding something hitherto unknown to us, human programmers cannot possibly tell the machine where to turn or which ways (heuristics) to follow: they don't know where one should go; only nature, so to speak, knows where the truth lies, and can lead the discoverer out of its maze. I shall examine these two features in the present and in the following section, respectively.

As for the former feature, the following three points should be considered.

i) First, researchers may aim at discovering either (a) just empirical laws (as the law of constant acceleration); or (b) empirical laws connected to and organized by theoretical laws, understood as purely mathematical formulas, whose non-observative terms don't purport to represent anything and lack a physical interpretation (as Maxwell equations might be understood); or (c) both empirical and mathematical laws, but embedded in a model or representation of reality (such as the field theory of electromagnetism, the kinetic theory of gases, the quantum model of energy, etc.).

Perhaps the distinction between (b) and (c) is not a sharp one, for seldom mathematical formulas are *pure* projections from the data and are devoid of *any* bit of interpretation.[45] Yet, it is intuitively clear and plausible, and there is no doubt that the most important discoveries belong to kind (c), while those of kind (b), even if achieved independently of a representative model, usually ask for and are soon supplied with one. For instance, while Plack's black-body equation was first devised as a pure uninterpreted projection from the empirical data (discovery of kind (b)), it soon got a physical interpretation by the quantum model of energy transmission (discovery of kind (c)). The importance of kind (c) discoveries is both historical (for their resonance and effects) and theoretical (for their systematic, heuristic and cognitive power).

Hence, by setting the discovery of models as PI's task, HHNT make at least a attempt to pursue discovery in its broadest extension. On the contrary, all the other programs are designed merely to discover regularities of kind (a) or (b), hence cannot be considered attempts to mechanize human discovering ability in its entirety.

ii) The second point concerning discovery as modelization is that even if we aim simply at laws as regularities, we must first decide which bodies of data we should consider and analyze in order to find relevant and interesting generalizations. For instance, by considering the values of pressure, volume and temperature in gases, we may find that they are linked by the ideal gas law. But how did we come to select exactly these parameters, to the exclusion, e.g., of smell, inflammability, and thousand others? Equally, by considering the periods and distances from the sun of planetary orbits, we may discover that they are significantly related (i.e., by Kepler's third law: the squares of times are proportional to the cubes of mean distances). But there is an endless number of different parameters one might have considered for similar relations.

In fact, the search for regularities is usually guided by models (provisional as they may be) the scientists entertain for the reality they are investigating. Even a vague idea of gas as consisting of particles subjected to mechanical forces will suggest that, e.g., volume and pressure may be relevant and mutually dependant, while smell is not. Equally, Kepler tried out a number of models as a guide to find a function for planetary distances from the sun: the model of planetary spheres nested into one another as respectively inscribed and circumscribed to the five regular solids; the model of a proportional relation between a planet's distance from the sun, and the density of the metal associated with it (mercury for Mercury, lead for Venus, etc.);[46] and more.

In many real life cases of discovery, the hardest problem was just to select the relevant parameters, for once these were selected, finding the relation holding among them proved rather straightforward. Now, neither Simon nor Turing's disciples say much on this question, for their programs may get to work only *after* being fed with the relevant data. For instance, Bacon$_3$ will apply its heuristics to whatever set of data about a gas it is given (e.g., mass, smell, specific weight, etc.); but only when fed with the right date will it find the ideal gas law.[47] Equally, only when fed with particular information concerning the primary structure of proteins has GOLEM been able to discover the desired law concerning their secondary structure.[48]

On the contrary, PI tries to select the relevant parameters by itself, and it can do so because it looks for analogical models (retrieving them in its background knowledge), and those models suggest the relevant parameters. For instance, in the example of the theory of sound, the analogies between sound on the one hand and water or ropes on the other suggest a wave model for sound, and such a model suggests frequency, wave-length, etc., as parameters to be considered for discovering the laws of sound.

iii) The third reason why models are necessary is that, as pointed out by Hanson,[49] theories and even laws are not a mere compendium of data: even once the relevant data are selected, the law or theory cannot always be straightforwardly extrapolated from them as if one plotted a curve upon a series of points in a Cartesian plane. For example, although all the relevant data coming from Brahe's observations were available to Kepler, he reached an apparently simple result as his first law (the planets' orbit is elliptical) only through long efforts, trials and errors.[50]

Now, it is well known that (a) for any body of data there are infinite possible laws or theories (principle of empirical underdetermination of theories); (b) there are no univocal criteria for choosing among laws or theories the most simple, or the most elegant, etc.; and (c) it may not always be easy to find a law or theory that is (by any criterion) simple or elegant.

Hence, if a law or theory were a merely abstract structure, with the only constraint procedures to follow, it couldn't be pursued by any computer program, and it would be of no concern to the epistemologists. In fact, this may be another reason why the logical positivists, who had an instrumentalist attitude toward theories and rejected Campbell's claims on the role of models and analogies in science,[51] showed little interest in the context of discovery and gave up the search for a logic of discovery.[52]

On the contrary, if a theory is a model representing an unknown piece of reality, it must not only be coherent with its data, but also *explain* them, and furthermore be coherent with the rest of our conceptions of reality. In this case, there is a complex but rational path of reasoning going back and forth from data to hypotheses, involving reasons, plausibility considerations, analogies, abductive inferences, assumptions, models, etc.: the complex but rational path exemplified by the accounts of great discoveries given by the historians or by the protagonists themselves.[53] This is more or less the lesson Aronson draws from the Flatland example: to a bidimensional being living on a plane suddenly appears a point, immediately enlarging to become a circle; the circle grows larger, then shrinks, becomes a point and disappears. If our being is an instrumentalist, sticking to his bidimensional data, he may find different formulas relating the changing dimension of the circle, its velocity, etc., none better than another and none explaining anything; moreover, he has no reason to link the phenomenon in significant generalizations with similar phenomena, like the following: a square appears, remains for a while, then disappears. Instead, if he is a realist, willing to entertain models of a third dimension, one hypothesis naturally links and explains all the features of one phenomenon, and different phenomena together: first, a sphere is passing through the plane, then, a cube is passing through it.[54]

If this is true, a realist interpretation of theories (viewing them as representations of reality) is presupposed by any account of the rationality of scientific discovery. By the same token, no instrumentalist computer (searching for laws or theories as pure projections from empirical data, as the programs written by Simon's group and in the Turing tradition) may become a machine for discovery. Only a realist program like PI, striving to find models of reality, has such a potentiality.[55]

(An anti-realist *à la* Van Fraassen may grant that models are necessary for heuristic purposes, but deny any need to believe them as true representations of reality.[56] Such an objection faces the same kind of reply facing Van Fraassen's position in general: how are we to explain the heuristic power of models, if they are not potentially true representations of reality? This reply may be challenged in turn, but the ensuing discussion cannot be pursued here).of being mathematically compatible with its data (i.e. of "saving the phenomena"), the process of discovery would be a sort of mysterious pulling the rabbit out of the hat: it would be a question of lucky intuition, there would be no rational

7. Discovery and Realism: the Role of Nature

If we grant that the task of a discovery machine is finding models, how does it select the *right* model? This has to do with the second feature enabling PI to be autonomous from human guidance: the fact of letting nature itself lead the program's search for discovery. This is achieved by building into the program two mechanisms: the first that, triggered by empirical inputs, through spreading activation produces an open number of models of reality; the second that, through the message competition and profit redistribution, is sensitive to nature's feedback: any step taking the system closer to its goal (producing a correct model of reality) activates a reward feedback, strengthening that move, and *vice versa* for steps in the opposite direction. If such mechanism is flexible and adaptable enough to match (to a certain extent, at least) the complexity and graduality of the environment, then the system should be able to lead to results that cannot be foreseen in advance, just as it happens with human procedures of discovery. The general model, here, is that of other well known adaptive and self-directed systems, such as the free market (an analogy explicitly exploited by HHNT), and natural selection.

Actually, both evolutionary studies and cognitive sciences seem to indicate that natural selection itself has built into the human mind a similar couple of mechanisms. Thus, it is also plausible that a program like PI is a good model of human abilities, including abilities to search and discover. Even this feature of HHNT's approach speaks for a realist interpretation of science: if the evolution and choice of models is constrained by the environment's feedback, then they do represent at least some aspects of reality. It is as if nature used human or mechanical researchers to paint its own self-portrait! Moreover, if the attitude of building models has been built into the human mind by natural selection, it is plausible that such models are true representations of reality, for acting on the basis of true representations warrants success in the search for food, defence from predation and reproduction.

Once again, anti-realists may object that (a) nature's feedback on the modelling processes of programs like PI are just empirical data: hence, such feedback simply warrants empirical adequacy, and (b) empirically accurate models shall be equally useful, for evolutionary purposes, as true models.[57] Still, we must ask *how* a living organism or a computer program can get to an empirically adequate model: the phenomenical aspect of nature is extremely complex, and there is no chance, either for living organisms or for machines, of building a general and detailed table of empirical data, yielding particular predictions for all kinds of situations, just by collecting empirical data and introducing *ad hoc* adjustments each time a prediction is not born out by experience. The only chance is to bet that underneath the great surface complexity there stand relatively simple ontologies and mechanisms that can be rationally understood, and try to model them; empirical predictions can then be deduced from the models. Now, if such a bet were not generally and approximately correct, it would be surprising that relatively simple models yielded arrays of predictions that are so complex, detailed, and largely born out by experience. Further, if such a bet were just an idle extrapolation, it would be surprising that natural selection had inbuilt it into the cognitive procedures of highly evoluted organisms; if instead it is the main road for discovery, it becomes clear why programs like PI have a chance of mechanizing discovery that earlier programs lacked.

References

Abbott, E. (1953) *Flatland*, Dover Press, New York.

Achinstein, P. (1980) "Discovery and Rule-Books" in Nickles, T. ed. (1980), 117-132.

Aronson, J.L. (1984) *A Realist Philosophy of Science*, The Macmillan Press, New York.

Bacon, F. (1620) *Novum Organum*.

Braithwaite, R.B. (1953) *Scientific Explanation: A Study of the Function of Theory, Probability and Law in Science*, Cambridge University Press, Cambridge Mass.

Braithwaite, R.B. (1962) "Models in The Empirical Sciences", in Nagel, E. ed., *Logic, Methodology and Philosophy of Science*, Stanford University Press, Stanford, Ca. 224-231.

Campbell, N.R. (1952^{1921}) *What is Science?* Dover, New York.

Campbell, N.R. (1957^{1919}), *Foundations of Science. The Philosophy of Theory and Experiment,* Dover, New York.

Curd, M. (1980) "Logic of Discovery: Three Approaches", in Nickles (1980), 201-219.

Fann, K.T. (1970) *Peirce's Theory of Abduction*, Nijhoff, the Hague.

Gentner, D, Holyoak, K.J., Kokinov, B.N. (2001) *The Analogical Mind: Perspectives from Cognitive Science*, MIT Press, Cambridge, Mass.

Gillies, D. (1996) *Artificial Intelligence and Scientific Method*, Oxford University Press, Oxford.

Hacking, I. (1983) *Representing and Intervening*, Cambridge University Press, Cambridge Mass.

Hanson, N.R., (1958) *Patterns of Discovery. An Inquiry into the Conceptual Foundations of Science*, Cambridge University Press, Cambridge Mass.

Hempel, C. (1965) *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, Free Press, New York.

Hempel, C. (1966) Philosophy of the Natural Sciences, Prentice-Hall, Englewood Cliffs, N.J.

Holland, J., Holyoak, K., Nisbett, R., Thagard, P. (1986) *Induction. Processes of Inference, Learning and Discovery* MIT Press, Cambridge, Mass., London, Engl.

Holyoak, K.J., Barnden, J.A. eds. (1994) *Analogical Connections*, Ablex, Norwood.

Holyoak, K., Thagard, P. (1995) *Mental Leaps*, MIT Press, Cambridge, Mass.

International Studies in the Philosophy of Science (1992) VI, 1.

Koyré, A. (1961) *La révolution astronomique*, Herman, Paris.

Kuhn, T. (1962) *The Structure of Scientific Revolutions*, The University of Chicago Press, Chicago, Ill.

Kuhn, T. (1970) "Reflections on my Critics", in Lakatos, I., Musgrave, A. eds. (1970).

Minds and Machines **14:** 21–42, 2004.

Lakatos, I. (1970) “Falsification and the Methodology of Scientific Research Programs”, in Lakatos, I., Musgrave, A. eds. (1970).

Lakatos, I., Musgrave, A. eds. (1970), *Criticism and the Growth of Knowledge*, Cambridge, University Press, Cambridge. Mass.

Langley, P., Simon, H., Bradshaw, G., Zytkow, J. (1987) *Scientific Discovery. Computational Explorations of Creative Processes*, MIT Press, Cambridge, Mass., London, England.

Laudan, L. (1980) “Why Was the Logic of Discovery Abandoned?”, in Nickles (1980), 173-183.

Magnani, L., Nersessian, N.J., Thagard, P. (eds.) (1999) *Model-based Reasoning in Scientific Discovery*, Kluwer Academic/Plenum Publishers, New York.

Mill, J.S. (1843) *A System of Logic, Ratiocinative and Inductive, Being a Connected View of the Principles of Evidence and Methods of Investigation*, Parker, London.

Molière, (Poquelin, J.B.) (1673) *Le malade imaginaire*.

Newton, I. (1687) *Naturalis Philosophiae Principia Mathematica*.

Nickles, T. (1980a) “Introductory Essay”, in Nickles, T. ed. (1980),1-60.

Nickles, T. ed. (1980b), *Scientific Discovery. Logic and Rationality*, Reidel, Dordrecht.

Pasquinelli, A. (1964) *Nuovi principi di epistemologia*, Feltrinelli, Milano.

Peirce, C.S. (1931-35) *Collected Papers of Charles Sanders Peirce*, I-VI (C. Harshorne and P. Weiss eds.), Harvard University Press, Cambridge, Mass.

Peirce, C.S. (1958) *Collected Papers of Charles Sanders Peirce*, VII-VIII (A. Burks ed.), Harvard University Press, Cambridge, Mass.

Pera, M. (1982) *Apologia del Metodo*, Laterza, Bari.

Pera, M. (1991) *Scienza e retorica*, Laterza, Bari.

Putnam, H. (1962) “The Analytic and the Synthetic”, in Feigl, H., Maxwell, G. eds., *Minnesota Studies in The Philosophy of Science* III, University of Minnesota Press, Minneapolis, 350-397.

Reichenbach, H. (1961) *Experience and Prediction: an analysis of the Foundations of Science*, University of Chicago Press, Chicago, Ill.

Searle, J. (1980) “Minds, Brains and Programs”, *Behavioral and Brain Sciences* III, 417-424.

Thagard, P. (1992) *Conceptual Revolutions*, Princeton University Press, Princeton, N.J.

Thagard, P. (1996) *Mind. Introduction to Cognitive Science*, MIT Press, Cambridge, Mass.

Thagard, P. (1998) *Computational Philosophy of Science*, MIT Press, Cambridge, Mass.

Turing, A. M. (1969) “Intelligent Machinery”, Report, National Physics Laboratory, in Meltzer, B., Michie, D. eds. *Machine Intelligence* V, Edinburgh University Press, Edinburgh 3-23, now in Ince, D.C. ed., (1992) *Collected works of A.M. Turing: Mechanical Intelligence*, North Holland, Amsterdam, 107-127.

Van Fraassen, B. (1980) *The Scientific Image*, Clarendon Press, Oxford.

[*] An earlier version of this paper was published online as "Artificial intelligence, logic of discovery and scientific realism", *Isonomia* 2002 (http://www.uniurb.it/Filosofia/isonomia/epistemologica.htm).

[1] See Hempel (1966) ch. II

[2] See ibidem.

[3] Mill (1843) III, ch. XI, § 3.

[4] See Reichenbach, (1961), pp. 350 ff.

[5] See for instance Peirce (1931-35) vol. VI, p.358; Peirce (1958) vol.VII, p.122; Fann (1970).

[6] Pera (1982), p.84; Laudan (1980), p.174; see also Curd (1980).

[7] O the question of scientific discovery, see the web sites http://server.math.nsc.ru/LBRT/logic/vityaev/; www.aaai.org/Press/Reports/Symposia/Spring/ss-95-03.html; http://www.unipv.it/webphilos_lab/courses/papers/creat_proces.htm

[8] This is usually called the *weak* Artificial Intelligence hypothesis, as contrasted with the *strong* hypothesis, namely that that machines can be built to perform the same tasks *by the same processes* as human intelligence. Actually, the original distinction between weak and strong Artificial Intelligence put forward in Searle (1980) was slightly different: according to the weak hypothesis machines can simulate thought, without actually having cognitive states, while according to the strong hypothesis machines really can think, understand, and have other cognitive states. The weak/strong contrapposition is also used in the literature to indicate different distinctions: machines can only work on data supplied by humans / machines can produce data themselves; machines can perform some of the intelligent tasks performed by humans / machines can perform all the intelligent tasks. See the following web sites:
http://www.comp.glam.ac.uk/pages/staff/efurse/Theology-of-Robots/Arguments-for-Strong-AI.html
http://sern.ucalgary.ca/courses/CPSC/533/W99/intro/tsld024.htm
http://www.faqs.org/faqs/ai-faq/general/part1/section-4.html
http://www.comp.glam.ac.uk/pages/staff/efurse/Theology-of-Robots/Arguments-Against.html
for a general introduction to Artificial Intelligence, see http://padova.fimmg.org/ring/docs/ai_tutorial.htm

[9] That is, supporters of the strong Artificial Intelligence hypothesis: see note 8.

[10] Ccerning computational work on scientific discovery, see the following web sites:
http://www.wam.umd.edu/~zben/Web/JournalPrint/printable.html
http://dmoz.org/Computers/Artificial_Intelligence/Creativity/Scientific_Discovery/
http://www.aaai.org/Pathfinder/html/discovery.html
http://directory.vaionline.it/Siti_Mondiali/Computers/Artificial_Intelligence/Creativity/Scientific_Discovery/
http://www.isle.org/~langley/discovery.html; http://www.citeseer.nj.nec.com/105475.html

[11] Langley, Simon, Bradshaw, Zytkow (1987).

[12] See Bacon (1620), Preface; I,61.

[13] Langley, Simon, Bradshaw, Zytkow (1987) p.76.

[14] Ibidem, pp. 88-108.

[15] Ibidem, p. 135.

[16] Ibidem, pp. 129-130.

[17] See Molière (1673), troisieme intermède, quoted in. Pasquinelli (1964), p.87.

[18] See Langley, Simon, Bradshaw, Zytkow (1987), pp.130-132.

[19] Ibidem, pp. 160-167.

[20] Ibidem, pp.170-185.

[21] Ibidem, pp.185-191.

[22] See ibidem, respectively chs. 6, 7 and 8.

[23] Ibidem, pp.248-251.

[24] See Kuhn (1962), chs. II, III, IV.

[25] See for instance the monographic issue *International Studies in The Philosophy of Science* (1992).

[26] In Gillies (1996), ch. 2.1.

[27] Gillies (1996), ch. 2 throughout.

[28] Ibidem, ch. 2.6.

[29] Ibidem, ch. 2.1.

[30] See for instance Aristotle, *Nichomachean Ethics*, I, 3, 1094b-1095a; Kuhn (1970), § 5: Irrationality and the Choice Among Theories; Pera (1982), pp.29 ff.; Pera (1991), pp.66-74, 117-124, passim.

[31] Holland, Holyoak, Nisbett, Thagard (1986).

[32] Characterized in Kuhn (1962).

[33] In the sense of Lakatos (1970).

[34] Ibidem, chs. 1, 2.

[35] See Putnam (1962), pp.378-379.

[36] See Holland, Holyoak, Nisbett, Thagard (1986), ch. 3.

[37] Notice that, as mentioned earlier, this inference has not generated new concepts, but only a new generalization involving already known concepts.

[38] Holland, Holyoak, Nisbett, Thagard (1986), p.45.

[39] Ibidem, ch. 10.

[40] See for instance Newton (1687) III, Regulae Philosophandi, rule 2.
[41] See Turing (1969). I owe this reference to Teresa Numerico.
[42] See Holland, Holyoak, Nisbett, Thagard (1986), pp.25-27.
[43] On this question see, Nickles (1980a), and Laudan (1980).
[44] See for instance Achinstein (1980).
[45] For instance, Planck worked out his black-body formula as complex equation approximating two different fuctions which accounted for empirical data respectively at low temperatures and high frequencies (the formula know as "Wien's law") and at high temperatures and low frequencies (a formula reflecting the data recently found by Ruben). He didn't (in the beginning) attach any physical meaning to the crucial discontinuity term of his formula, nor associated it to a particular (quantum) model of energy transmission. Yet, his formula included terms, as frequency, already embedded into a representative model (the wave model of energy, with crests, troughts, and frequency as the number of crests per time unit).
[46] Respectively in *Mysterium Cosmographicum* and in *Epitome Astronomiae Copernicanae*: see Koyré (1961), Part II, sections 1, 3.
[47] See Langley, Simon, Bradshaw, Zytkow (1987), ch.3.
[48] See Gillies, (1996), § 2.6.
[49] In Hanson (1958), ch. 4.
[50] As he explains in his *Astronomia Nova*. See also A. Koyré (1961), Part II, section 2.
[51] See Campbell, (1952), p.96; Campbell (1957), pp. 123 ff.
[52] See Hempel (1965), pp. 444 ff.; Braithwaite (1962), pp.230-231; Braithwaite (1953), p.51.
[53] As for instance Kepler in *Astronomia Nova*.
[54] See Aronson (1984), ch 7, § 2. The idea comes from Abbott (1953).
[55] O the role of models in scientific discovery, see the web sites: http://lgxserver.uniba.it/lei/recensioni/crono/2000-06/magnani.htm; http://philos.unipv.it/courses/inf_proc.htm
[56] See Van Fraassen (1980), ch.2, § 1.3.
[57] Similar objections are raised, for instance, in B. Van Fraassen (1980), ch.2, §§ 3, 4, 7, and Hacking (1983), ch. 3.