# The Hierarchies of Knowledge and the Mathematics of Discovery

CLARK GLYMOUR[1]
*Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, U.S.A.*

**Abstract.** Rather than attempting to characterize a relation of confirmation between evidence and theory, epistemology might better consider which methods of forming conjectures from evidence, or of altering beliefs in the light of evidence, are most reliable for getting to the truth. A logical framework for such a study was constructed in the early 1960s by E. Mark Gold and Hilary Putnam. This essay describes some of the results that have been obtained in that framework and their significance for philosophy of science, artificial intelligence, and for normative epistemology when truth is relative.

**Key words.** Learning, induction, logic, relativism, epistemology.

Philosophy before and after mid-century is united in a rejection of a central goal of traditional epistemology from Plato to Boole: a theory of discovery. Plato and Aristotle thought the goal of philosophy, among other goals, was to provide methods for coming to have knowledge. This same conception utterly dominated philosophy in the 17th century. It was Descartes' claim to have found such a method, and the disputes between him and his critics were in part over what it is to be a method of discovery at all. Leibniz not only advanced the conception of method, but provided a thesis about it that guided logical investigations into the 20th century. In my view, the central 18th century dispute in philosophy, between Hume and Kant, was fundamentally about whether we can have methods of inquiry that can be known to be reliable. The latter part of the century provided in Richard Price's interpretation of Bayes' probabilism yet another proposal for a universal method of discovery. English-speaking philosophers of the succeeding century were equally absorbed with discovery: John Stuart Mill popularized a method plagarized from Bacon and, in aid of a method for discovering causal relations from probabilities, George Boole made the largest advance in logical theory since Aristotle.

But after 1925 or thereabouts, there was in philosophy almost nothing more of methods of discovery. A tradition that joined together much of the classical philosophical literature simply vanished. From about 1930 to about 1960 philosophy of science was in fashion, and certain questions of epistemology – the existence of sense-data, for example, and the role of stipulations in our systems of belief – won the attention of even the most eminent philosophers. These were not the sorts of epistemological questions, however, that were the principal focus of epistemology for major philosophical writers before our century. And since the middle of the 1960s scarcely any major philosopher has thought even these

*Minds and Machines* 1: 75–95, 1991.

epistemological questions worth much bother, let alone questions as to the best method of making discoveries or the limits of the discoverable. The latter questions are now commonly thought to be absurd and to make false and naive presuppositions of one kind or another. As late as the 1980s a philosophical reporter could truly announce that most philosophers hold that there is and can be "no systematic useful study of theory construction or discovery."[2] (Insofar as they gave any heed to the question at all, the same might well have been said of most scientific practitioners: of statisticians, social scientists, economists, physicists.) The pre-eminent view among philosophers nowadays is that claims to knowledge, or to the possession of normative standards for methods of acquiring knowledge, are so much rhetoric, so much politics; truth, insofar as it is a useful notion at all, is relative to the conditions of the believer, and there are no matters of fact independent of the inquirer and the community.

In contrast, traditional epistemological questions were at the very heart of this century's revolutionary developments in logic and computation theory. From the mathematical logic of Hilbert, Gödel and others, from the theory of computation created by Church, Post and Turing, and from the theory of recursion there developed in the last twenty five years a beautiful mathematical theory of methods of discovery and of the limits of knowledge, a theory that directly addresses the central epistemological concerns of the great philosophical tradition before this century. It is a theory about discovery that is nice in itself, of use to serious scientific concerns, and even applies to the concerns of the effete: it contains epistemological norms for those who hold that truth is relative to conceptual scheme. The subject has lain almost completely hidden from the view of philosophers and practitioners. I did not come upon it until ten years ago, after I had written a book on epistemology that concluded by calling for the creation of a theory whose fundamentals had already existed for fifteen years. My aim is to tell you something about the development of this subject, and to discuss some of its applications.

## The Platonic-Positivist Epistemic Hierarchy

Plato's *Meno* presents a view about inquiry and discovery that has had an enduring appeal. In that dialogue the Socratic task is to learn truths of a special kind. From a logical point of view, what is to be learned, for example about the nature of virtue, is a universal biconditional sentence without disjunction that can serve as an appropriate definition, e.g. of "is virtuous." The learning is by example and counterexample. Socrates presents examples of virtuous things and their features, and examples of things that are not virtuous and their features; the correctness of the data of the examples and counterexamples is never in doubt. What is it that Plato requires in order for someone to have discovered in this way the answer to the question, "What is virtue?" To know the answer, one must hit upon the correct definition of virtue, and *know* that one has done so. One must have the kind of certainty that amounts to a dogmatism, and reserves no right to alteration: *opinion* can change, knowledge cannot. How such knowledge is possible is the point of Meno's challenge to Socrates: How will Socrates recognize the truth when he comes upon it? Plato's answer appeals to an internal oracle that somehow guarantees the correctness of certain definitions.[3] Without the oracle, nothing is firm save the examples and the counterexamples.

In the 1930s, philosophical conceptions of discovery were essentially Plato's but without the oracle. It was supposed that there are some matters that are simply data, and either permanently or contextually fixed. They are the "sense data" or "observation statements" or "protocol sentences." They met the Platonic criterion for the discoverable: once accepted in the context of some inquiry, one could be sure that they would not be abandoned. Only two other kinds of discoveries met that criterion: mathematical truths, and sentences verified by the data. With only a little logical knowledge, philosophers in this period understood the verifiable and the refutable to have special logical forms, namely as existential and universal sentences respectively. There was, implicitly, a positivist hierarchy (see Figure 1). Positivists such as Schlick confined science and meaning to singular data and verifiable sentences; "anti-positivists," notably Popper, confined science to the singular data and falsifiable sentences. In both cases, what could be known or discovered consisted of the singular data and verifiable sentences, although there is a hint of something else in Popper's view. In Popper's conception of inquiry consists of conjecturing falsifiable sentences and attempting to falsify them; Popper in effect agreed with Plato that knowledge requires a kind of unalterability, but unlike Plato he did not think that the process of science obtains knowledge. Popper and the positivists agreed that there could not, in any case, be an *algorithm* for carrying out scientific inquiry. Why not?

In the Platonic conception, an algorithm for scientific discovery must be a procedure that examines data and, after a finite time, announces the truth. Whenever the procedure results in such an announcement, it must be correct. There must be no possibility of revision. The Platonic conception of an algorithm for discovery was also the philosophers' conception in the 20th century, but the

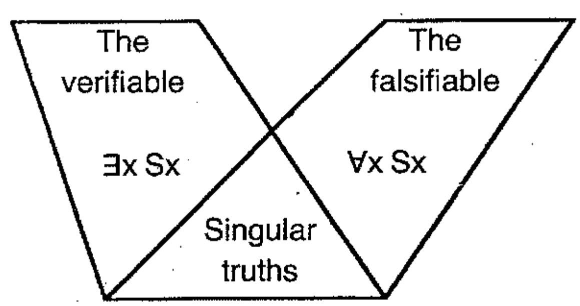



Fig. 1. The positivist hierarchy.

philosophers did not hold with oracles. They rightly believed that no such algorithm is possible for the claims of science. There are algorithms of this kind that will reliably conclude when universally quantified formulas are false, but none that will conclude when they are true, and science is filled with what appear to be universal claims. That is one reason for denying that there is an algorithm for scientific discovery. Another is the changes in physics that had taken place in the three decades before 1930, and that were familiar to many philosophers. If there were an algorithm for discovery, one could only think that the practice of science embodied a social version of it. And if there were such an algorithm embodied in scientific practice, then most certainly by the latter half of the 19th century that algorithm had announced the truth of Newtonian physics. John Tindall, for example, announced in popular lectures that the framework of physics was fixed forever; all that remained to do was to find the various force laws. The years between 1905 and 1926 utterly demolished Tindall's claim. A further reason for thinking that empirical discovery could be subject to algorithm was the authority of Albert Einstein, who with charming inconsistency claimed both determinism in human affairs and that scientific theories are "free creations." For Popper – who quite confused a psychological question with a mathematical issue – it sufficed to quote Einstein to disprove the possibility of a discovery algorithm; for Carnap it sufficed to quote Popper quoting Einstein.

## The Entscheidungsproblem and Algorithms for Mathematical Discovery

Frege's remarkable logical achievement was a theory of proof; a proof theory for what is now known as first-order logic is explicit in the *Begriffschrift*. Two questions naturally arose for which Frege provided no answer: Is the system of proof complete? Is there an algorithmic procedure that will, for any formula, decide whether or not it is provable? The importance of these questions is, of course, epistemological. Hilbert and others suggested that a positive answer to the two questions would show that it is in a sense in principle possible to carry out Leibniz's vision. If Hilbert and Ackermann's proof theory, for example, were complete and admitted a decision procedure, then there would exist a method to discover the consequences of any first order axiomatization.

Not long after the questions had been clearly formulated, Gödel answered the first question affirmatively, and gave reason to think the answer to the second question is negative. Church and Turing settled the question altogether. Insofar as the philosophical community took note of the epistemological significance of these results, they cemented the conviction that there can be no such thing as algorithms for empirical discovery, and no interesting theory about them. And yet from a logical point of view, because of Gödel's completeness theorem, the undecidability of the validity of first order formulas did not quite kill the idea of an algorithm for mathematical discovery. Rather, it throws into clear relief an epistemological idea about what it is to come to know something, an idea that is quite different from the Platonic and the Positivist conception.

Consider trying to discover whether or not a certain first-order formula $\Phi$ is valid. Since Hilbert and Ackermann's system is complete, if $\Phi$ is valid there is a proof of it. Since there is a decision procedure that decides whether or not a finite sequence of formulas is a proof, and since the collection of all finite sequences of well-formed formulas can be effectively enumerated, we can imagine a procedure that examines each finite sequence of formulas in such an enumeration in turn, and checks to see whether or not it is a proof of $\Phi$, and stops when a proof is found. Call this procedure $P$. The procedure $P$ will eventually find a proof if in fact $\Phi$ is valid. Otherwise the procedure will continue on forever. Suppose now that we adopt the following rule for formulating hypotheses as to whether or not $\Phi$ is valid:

If at stage $n$, $P$ does not say that a proof of $\Phi$ has been found, conjecture that $\Phi$ is not valid.

Is this an algorithm for acquiring knowledge about logical truth? Clearly not, if your conception of what it is to know is Plato's. Using this algorithm, if $\Phi$ is not valid, there is no time at which you can be certain of that fact, no time at which you can rock back and say, "No further evidence is needed." But the rule for formulating conjectures has a property that suggests a different conception of what it is to know: *Using this rule, there is some finite stage after which your conjectures as to whether or not $\Phi$ is valid will always be correct.* Eventually you will be right forever after, although if $\Phi$ is not valid you will never know when that stage has arrived, and you will never be able to dispense with further evidence. Perhaps that is all *knowledge* requires. Perhaps you know the truth about the validity of $\Phi$ if you are disposed to conjecture by a rule that has this convergence property and you have in fact reached a stage after which conjectures made according to that disposition are always correct. Call this sort of relation *knowledge in the limit.*

I doubt that there is one true account of what it is to know, but certainly this is an interesting knowledge relation, and one we can have even when we can't have the sort of knowledge Plato required. When can we have knowledge in the limit, and when not? We have just seen that we can have it for the validity of any first order sentence. When can we have it for empirical issues? There's a good question.

## Turing, Putnam and Gold

The epistemological idea about knowledge in the limit is implicit in many contexts in the 20th century. Abraham Robinson remarked that something like it is to be found in Gödel's proof of the completeness theorem. But the articulation of the idea came almost simultaneously in the 1960s from two independent sources, Hilary Putnam and E. Mark Gold. It seems likely that Putnam took the idea from Hans Reichenbach and combined it with reflections on Turing's conventions for the output of a computing machine. In Putnam's words:

we know what sets are "decidable" – namely, the recursive sets (according to Church's Thesis). But what happens if we modify the notion of a decision procedure by (1) allowing the procedure to "change its mind" any finite number of times (in terms of Turing machines: we visualize the machine as being given an integer (or an $n$-tuple of integers) as input. The machine then "prints out" a finite sequence of "yesses" and "nos." The *last* "yes" or "no" is always to be the correct answer.); and (2) we give up the requirement that it be possible to tell (effectively) if the computation has terminated. I.e., if the machine has most recently printed "yes" then we know that the integer put in as input must be in the set *unless the machine is going to change its mind*; but we have no procedure for telling whether the machine will change its mind or not.

The sets for which there exist procedures in this widened sense are decidable by "empirical" means – for, if we always "posit" that the most recently generated answer is correct, we will make a finite number of mistakes, but we will eventually get the correct answer. (Note however, that even if we have gotten to the correct answer (the end of the finite sequence) we are never *sure* that we have the correct answer.

Instead of requiring that the sequence of "yesses" and "nos" be finite and non-empty, we may require that it should always be infinite, but that it should consist entirely of "yesses" (or entirely of "nos") from a certain point on; the class of predicates obtained. . .is easily seen to be unchanged.

Gold called such sets "limiting recursive", Putnam called them the extensions of "trial and error predicates." Gold's terminology has stuck. Gold and Putnam each proved the same main theorem: A set is limiting recursive if and only if it is in $\Delta_2$ in the arithmetic hierarchy. Gold proved a similar result for recursive functionals. Putnam's proof is easy and instructive.

Recall that the $\Delta_2$ sets in the arithmetical hierarchy are the following: A set $S$ is $\Sigma_2$ if there is a formula $\exists x \forall y R(xyz)$ such that $R$ is a recursive predicate of triples of numbers and $S$ is the set of all numbers satisfying the formula. A set is $\Pi_2$ if its negation is $\Sigma_2$. If you drive the negation through the quantifiers in a $\Sigma_2$ formula you get a formula that is universal existential with a recursive predicate. A set is $\Delta_2$ provided that it is both $\Sigma_2$ and $\Pi_2$. In other words, a set is $\Delta_2$ provided that there is a formula $\exists x \forall y R(xyz)$ such that $R$ is a recursive predicate of triples of numbers and $S$ is the set of all numbers satisfying the formula, and also there is a formula $\exists x \forall y P(xyz)$ such that $P$ is a recursive predicate of triples of numbers and the complement of $S$ is the set of all numbers satisfying that formula.

Suppose that $S$ is an arbitrary set of numbers, and there is a Turing machine $T$ that for every number $n$ coverages in the limit to "yes" if $n$ is in $S$ and converges to "no" if $n$ is in the complement of $S$. The predicates '$T$ on input $x$ converges in the limit to "yes;"' and '$T$ on input $x$ converges in the limit to "no"' can each be formalized in number theory, e.g. $\exists m \forall n\ n > m \rightarrow T(x, n) = 1$ and $\exists m \forall n\ n > m \rightarrow T(x, n) = 0$, where '$T(x, n)$' denotes a total recursive function. So they are each in $\Sigma_2$. Since by assumption $T$ must for any input converge to "yes" or "no" and cannot forever vacillate, '$T$ on input $x$ converges to 1' is satisfied by a value of $x$ if and only if for every stage $y$ of computation such that $T(x, y)$ is not 1, there is some later stage $z$ for which $T(x, z) = 1$. So '$T$ on input $x$ converges in the limit to "yes" is also equivalent to $\forall y \exists z[T(x, y) \neq 1 \rightarrow (z > y)\ \&\ T(x, z) = 1)]$. So the predicate is also $\Pi_2$. Hence $S$ is a $\Delta_2$ set.

Suppose, conversely, that $S$ is a $\Delta_2$ set. Then $S$ is a set of numbers that satisfy $\exists x \forall y R(xyz)$ for some recursive $R$ and the complement of $S$ is the set of numbers that satisfy $\exists x \forall y P(xyz)$ for some recursive $P$. Let $T1$ be a Turing machine that computes (in the usual way) $R(xyz)$ and let $T0$ be a Turing machine that computes (in the usual way) $P(xyz)$. Given input $z$, the set of all triples $xyz$ can be effectively enumerated. Let $\langle n, m, z \rangle_i$ denote the $i$th triple in some such enumeration. For each $h$ using $T1$ a machine $T11$ can check effectively whether or not $\exists x \forall y R(xyz)$ is true in the set of all triples $\langle n, m, z \rangle_h$ for $i \geq h$. Let the output of $T11$ be 1 for the $i$th set of triples if $\exists x \forall y R(xyz)$ is thus satisfied and 0 otherwise. Similarly, using $T0$ a machine $T00$ can check effectively whether or not $\exists x \forall y P(xyz)$ is true in the set of all triples $\langle n, m, z \rangle_h$ for $i \geq h$. Let the output of $T00$ for the $i$th triple be 1 if $\exists x \forall y P(xyz)$ is thus satisfied and 0 otherwise. Now let $T(z, n)$ be the machine that on input $z$ gives an infinite string of outputs whose $n$th, $T(z, n)$, is 1 if $T11(n)$ is 1 and is preceeded by a longer uninterrupted string of 1s than is $T00(n)$, and let $T(z, n)$ be 0 otherwise. $T(z, n)$ is the machine that computes $S$ in the limit.

Although tied to computation, the idea behind Putnam's proof has a more general epistemological significance. Suppose given any triple of objects $\langle u, v, w \rangle$ you have some way of determining whether or not they satisfy $R(xyz)$. Never mind about computers, just some way. Suppose, over some domain you can investigate each triple of objects making the determination as you go. Then if $\exists x \forall y R(xyz)$ is true, you can know in the limit that it is: just keep guessing "yes" if the formula is satisfied for all triples (with $z$) you have seen so far, and "no" otherwise. If the formula is true after a finite time you will find a value of $x$ that in fact stands in the relation $R(xyz)$ for all values of $y$, and you will be correct in your guess ever after; if the formula is false, you will either converge to "no" or change from "yes" to "no" or back again infinitely often. And if it is the case that if formula $F$ is true you can know it in the limit, and also that if $-F$ is true you can know it in the limit, then by running the two inquiries jointly you can know in the limit whether or not $F$ (and likewise whether or not $-F$) is true. It looks as though what you can know in the limit is characterized by existential and universal quantification over what you can know in the Platonic way.

## Confirmation Relations and Languages

How does one get from the characterization of the limiting recursive sets of numbers to an understanding of empirical questions for which discovery methods do and do not exist? There was a direct route, which was not taken. Hilary Putnam seems to have come to the idea through two prior papers about limitations on the reliability of Carnapian confirmation functions.[5] His arguments assumed in effect that there is a collection of possible relational structures, and the learning procedure is given, singular fact by singular fact, the diagram of some structure in the collection. At each stage the learner must either guess a hypothesis or alter the probabilities it assigns to the hypotheses in light of the evidence. The question is whether the machine can eventually output the truth, or eventually always give the true hypothesis: a probability (or degree of confirmation) greater than 1/2. These papers are wonderfully prescient in seeing that confirmation theories are cogs in possible learning algorithms and in struggling to form a framework in which to evaluate such algorithms. They were unfortunately wrong in their optimism. Writing in 1963, Putnam saw that there was a rich structure to investigate and assumed that logicians and philosophers of science would turn to uncovering it. By and large save for his own work and Gold's that didn't happen, and by the time Putnam's vision was realized, confirmation theory no longer interested philosophers.

Gold applied the idea of limiting recursion to issues motivated by Chomsky's work rather than by Carnap's: the problems of language learning. The application is quite natural. Chomsky was concerned with Universal Grammar – the grammatical features shared by all possible human natural languages – and a principal constraint on that hypothetical grammar was that, whatever the set of possible human natural languages might be, it must be possible for a human to learn to parse any language in that collection. What collections of languages meet that condition?

Gold reformulated the question this way. Give the well-formed sentences of a language Gödel numbers. Then, syntactically, a language $L$ can be represented as a recursive set of numbers. One way to view a parser for the language is then as a Turing machine or other program that decides for any number whether or not it is the number of a grammatical string in the language. We can effectively enumerate the Turing machines, giving each program a number or index. Learning to parse a language implies that one has identified, at least implicitly, the index of a program for deciding the set of well-formed strings of that language. Suppose that the would-be learner receives the well-formed strings of the language in some order and never receives (or ignores) strings that are not in the language. Every string in the language eventually occurs, and a string may occur any number of times, even infinitely often. Suppose after each string is received the learner guesses a program (or an index for a program) that he conjectures will parse exactly the unknown language. For what collections of languages does there exist a learner who, no matter which language is the correct one and no matter in what order the data are received, will obtain limiting knowledge of the index of a program to parse the language?

Gold showed that there are simple collections of languages that cannot be learned in the limit by any possible learner, not even by one free of computational constraints. A famous and simple example is the collection consisting of all finite subsets of $N$ together with $N$.

Gold's paper was followed in the next twenty years by a great deal of work on language learning. The assumptions about data and convergence criteria were altered in various ways, notions of approximation introduced, relations among the paradigms were studied extensively, the effects of methodological strictures on the capacities of learners were studied, and ever more psychologically realistic learning constraints were investigated. Many of these results are presented in Osherson, Stob and Weinstein's *Systems That Learn*. One of the fundamental results of this literature was obtained by Dana Angluin, who provided a characterization of necessary and sufficient conditions for any subset of the collection of recursively enumerable languages to admit a learner that could identify any language in the collection in the limit no matter the order in which the strings of the language were presented as data. Of course these collections of alternative languages were necessarily countable.

## Learning Theories

Despite the interesting methodological structure of the studies of language learning in the limit, it was not evident just how to make it apply to the questions with which we began concerning methods of empirical discovery. The movement back to Putnam's original concerns began with Angluin's student, Ehud Shapiro.[6] Recall our discussion of the problem of deciding validity, and the existence of procedures that will decide validity in the limit. In the same way, there are procedures that will decide entailment in the limit. This suggests a sort of Popperian approach to discovery: formulate a hypothesis, gather evidence in the form of singular sentences and see if, in the limit, all of the evidence can be deduced from the hypothesis and no denial of any evidence sentence can be so deduced. Somehow order the possible hypotheses so that their testing, gathering further evidence, changing conjectures appropriately, etc., can be dovetailed. Shapiro described algorithms of this sort that find a true finite axiomatization of all of the atomic sentences true in a structure when such an axiomatization exists. The predicates occurring in the hypotheses must be the same as those occurring in the evidence.

Suppose we consider a collection of relational structures for a language. Imagine that one of the structures, we know not which, characterizes our actual circumstances. Whichever world is actual, we will receive from it a sequence of singular facts characterizing the diagram of the structure. The order of the sequence of data is not subject to our control. Generally we want something other than a true finite axiomatization that entails all of the true atomic sentences. What might that be?

It might be that we want to know which theory within a certain class of alternative theories is correct. Suppose so. We could learn a theory in the limit in at least two different senses. In one sense, called *EA* or uniform learning, we learn a theory by converging in the limit to a conjecture for that theory (or if the theory is not finitely axiomatizable and we insist that the outputs of our conjecturing process be finite objects, to a program for computing a set of axioms for the theory). So there exists a point after which all of our conjectures about the identity of the true theory are correct. In another sense, called *AE* or non-uniform learning, we could learn a theory by converging in the limit piece by piece. That is, for every theorem of the theory there exists a point after which every theory conjectured entails that theorem, and for every sentence that is not a theorem of the theory there exists a point after which no theory conjectured entails that sentence. Kevin Kelly and I characterized by syntactic classes the cases for first order theories in which the true alternative can (and cannot) be identified in the *EA* or *AE* sense, either by Turing computable learners or by learners that embody arbitrary functions – learners who have powers that transcend the computable. Later work extended the classification for *AE* theory learning to cases in which quantified sentences occur in the data.[7]

Another thing we might want in empirical inquiry is the answer to a specific question. We might consider discovery problems set up closer to those Putnam envisaged, in which a question is posed by a first order sentence whose truth value is to be determined, data is obtained from an unknown structure in a collection of alternative structures, and conjectures are made as to the truth or falsity of the sentence in the unknown structure. This case was investigated by Dan Osherson and Scott Weinstein,[8] who distinguished a number of alternative senses of convergence to the truth: the learner can converge to the correct truth value for $\Phi$ if $\Phi$ is true but possibly fail to converge otherwise; converge to the correct truth value for $\Phi$ if $\Phi$ is true but possibly fail to converge otherwise; converge to the correct truth value if $\Phi$ is false; or do both. They showed that for learners with "free will" *AE* theory learning is possible if and only if there is a learner who can converge to the truth for any sentence in the language of the theory. Importing methods from the investigation of language learning, they showed that various methodological principles, such as consistency and conservatism, restrict the scope of the reliability of any Turing-computable learner. And, by the same means, they characterized the conditions for which it is possible to obtain knowledge in the limit about the truth of a sentence, provided the number of alternative relational structures is countable and the learner is required to succeed on all possible orderings of the complete data true in a structure.

This work on learning theories had two obvious points of weakness, shared in principle by the work on language learning. First, the collections of alternative structures or theories over which discovery is possible may be *uncountable*. There are, for example, uncountably many distinct purely universal theories, but there is a learner (in fact a Turing computable learner) that will learn (*AE*) any purely universal theory. But the characterizations of necessary and sufficient conditions for knowledge in the limit, whether of languages or of the truth or falsity of first-order formulas, were restricted to cases in which the number of alternative structures is countable. Second, all of the investigations considered, whether of language learning or theory learning, assumed that every possible ordering of the data could occur. The learner is never permitted to have prior knowledge restricting the order in which the data arrive. In fact that is quite implausible both for language learning and for theory learning. To fully understand knowledge in the limit, these two artificial restrictions needed to be removed. Recent work by Kelly has removed them by returning to the original ideas in Gold's and Putnam's papers.

## The Hierarchies

Suppose that the facts that may occur as data, whether the strings in a language or singular or quantified formulas satisfied in a structure, are encoded as numbers. Then an infinite data sequence is an $\omega$ sequence of numbers. So a data sequence can be thought of as a function from $\omega$ to $\omega$, assigning to each finite ordinal the datum that occurs there. Now consider the set $B$ of all such sequences. Extensionally, any property of data sequences is a subset of $B$. For example, if the data are from languages, then for any language there is a subset of $B$ corresponding to the set of all infinite sequences of strings from the language; if the data are from a relational structure, then for any structure there is a subset of $B$ corresponding to the set of all infinite sequences of singular data. This suggests that the way to avoid the limitation of previous investigations to circumstances in which the data from a language or structure can occur in any order is to investigate which properties of data sequences can be known in the limit. That is, instead of thinking in terms of identifying languages or identifying relational structures or learning theories, let us ask the more general question: when can we know in the limit that the data sequence we are investigating has a specific property? If we know the answer to that question, the answers to other questions will follow as special cases.

Gold's and Putnam's papers suggest the following idea: what you can know to be true in the limit is what you can get by quantifying existentially-universally over what you can know Platonically. Some analogies transform this suggestion into guides for investigating the possibility of knowing properties of data sequences in the limit: What you can know Platonically is just the initial segments of data sequences. Quantifying existentially is analogous to taking infinite disjunctions, which is analogous to taking infinite unions. Quantifying universally is analogous to taking infinite conjunctions, which is analogous to taking infinite intersections. We can describe a hierarchy of collections of subsets of $B$ using an analogy. Consider any subset $S$ of $B$ for which there is a set of initial segments such that all and only data sequences in $B$ having those initial segments are in $S$. Call the collection of such subsets $\Sigma 1$. Consider the collection of subsets of $B$ each member of which is the complement of some set in $\Sigma 1$. Call this collection $\Pi 1$. Let $\Delta 1$ be the intersection of $\Sigma 1$ and $\Pi 1$. Now consider the collection of all subsets of $B$ that are (countable) unions of sets in $\Pi 1$. Call this collection $\Sigma 2$. Consider the collection of all subsets of $B$ that are (countable) intersections of sets in $\Sigma 1$. Call this collection $\Pi 2$. Let $\Delta 2$ be the intersection of $\Sigma 2$ and $\Pi 2$. Continue in this way forever and ever again. The result is a hierarchy that is closed upwards, the Borel hierarchy.

The sets in $\Sigma 1$ correspond to those properties such that if a sequence has such a property some computationally unbounded learner can eventually have Platonic knowledge that it does. (Just wait until one of the initial segments characteristic of the set appears.) The sets in $\Pi 1$ correspond to the properties such that if a sequence fails to have the property, some computationally unbounded learner can eventually have Platonic knowledge that it fails to have the property, i.e., that it has the complementary property. The properties such that some unbounded learner can have Platonic knowledge of whether or not the data sequence under investigation has that property are given extensionally by sets in $\Delta 1$.

We might suspect that the knowledge in the limit available to a computationally unbounded learner corresponds to sets in $\Delta 2$ in the Borel hierarchy. That is what Kelly proved. In fact he proved something stronger. We can think of "background knowledge" as given by a subset $K$ of $B$. Starting not with the sets of data sequences that share an initial segment but instead with intersections of such sets with $K$, we can build a relativized Borel hierarchy. Kelly has shown that if $P$ is any subset of $B$, a computationally bounded learner with background knowledge

$K$ can know in the limit whether or not a data sequence is in $P \cap K$ if and only if $P \cap K$ is in $\Delta 2$ in the hierarchy relativized to $K$.

And what if discovery must be done by computationally bounded agents? The key to the solution to that question lies in Gold's use of the recursive functionals. Consider a Turing machine learner at work on a sequence from $B$. There is actually some $\omega$ sequence the learner is receiving as data, and at each stage the learner outputs either 1 or 0. So the learner can be thought of as a partial recursive functional $T[t, n]$ where $t$ is the infinite sequence, and hence really a function from $\omega$ to $\omega$, and $n$ is the stage of data presentation. A Turing machine interpretation of such a functional is as a machine that can, for any $t$ and $n$, receive the first $n$ values of $t$ before producing an output. We are asking, in effect, not which sets of numbers are computable in the limit, but which sets of functions from the natural numbers to the natural numbers are computable in the limit.

Now just as there is a recursion theoretic (arithmetic) hierarchy for sets of numbers, there is a recursion theoretic (arithmetic) hierarchy for sets of functionals. A functional of type $\langle k, j \rangle$, is just a finite sequence of $k$ functions from $\omega$ to $\omega$ and $j$ numbers. A relation is a set of functionals all of the same type. We can think of relations of type $\langle 1, 0 \rangle$ as subsets of $B$. The recursion theoretic hierarchy can be constructed analogously to the Borel hierarchy, but using quantifiers rather than unions and intersections. In the same way, starting with background knowledge $K$, one can construct a relativized hierarchy. Kelly proved that a Turing computable learner with background knowledge $K$ can know in the limit whether or not a relation obtains if and only if the relation is $\Delta 2$ in this hierarchy. The same result is implicit in Gold's Theorem 4.

Together these results yield general characterizations for Turing computable and for computationally unbounded learners both of language learning in the limit and of detecting the truth or falsity of a first order formula in the limit. The characterizations are not limited to cases in which the number of alternative structures or languages is countable, and they do not require that one assume that every ordering of the data is possible.

Kelly's results don't close the subject; they open it up for application and investigation. For example, Kelly derives a characterization of conditions under which the truth or falsity of a given first order hypothesis can be known in the limit; one of the surprising consequences is that if any such problem can be solved by a computationally unbounded learner, it can also be solved by a Turing computable learner. So far as deciding in the limit the truth or falsity of a given first-order sentence, the Turing computability of the learner is no handicap. But that is not so when we consider the $AE$ learning of theories. We have only limited knowledge of when theories can be learned $AE$ by a Turing computable learner. We will come across a number of other open questions in what follows.

## Relativism

Whenever something is a lot of work folks are bound to look for reasons why it isn't worth the effort. There are lots of complaints about limiting analyses of learning that seem mere excuses. For example, that no one cares what happens in the long run. The reply is twofold: first, that short-run results are much to be desired but require strong background knowledge that we often fail to have, and, second, if you can't know the truth in the long run, you can't know it in the short run either. Another is that the results and techniques can't be applied to "real" – meaning other people's – problems. We'll see that they can be indeed. Still other excuses appeal to some relevant factor of inquiry that is not explicitly represented in the formal representations – experimentation, for example. The response is that it is in principle straightforward to include experimentation in the framework, and work to that end is already under way.

But there is a further objection that is more fundamental and that is surprisingly interesting. Suppose one denies, with many prominent contemporary philosophers, that there is any one common world of inquiry. Suppose one denies that there are any facts of experience to serve as data that are independent of the inquirer. Instead one holds that, depending on what one believes, on one's history, on the community to which one belongs, or other factors, there will be different data. Even suppose that depending on such factors the very character of logic may change. Then all of the results I have so far described are otiose; they do not apply, they are "inoperative."

These are the suppositions that dominate contemporary philosophical discussion. Their champions conclude that there are no such things as epistemological norms, because there are no such things as intelligible epistemological goals. I find these views enormously distasteful. Each time I read or hear some plump and comfortable academic saying such things I am overcome by images from *Darkness at Noon*. But that is no reason not to think about the epistemology of relative truth. It turns out to have an astonishing and intricate structure, altogether unseen and unexplored either by its advocates or its critics.

Suppose that the world of experience is a function of some feature of the inquirer. Even the most radical critics of science rarely hold that what one experiences depends on what one believes or does and on nothing else. So the world of experience is a function of features of the inquirer and of features, we know not what, that are not subject to the inquirer's power. For brevity let us call the first set of factor's the "conceptual scheme" of the inquirer and the second set of factors the "world in itself." Then the world of experience is a function of conceptual scheme, which is subject to the inquirer's choice and decision, and of the world in itself, which is not. We can think of the world in itself abstractly as simply a function that for each possible conceptual scheme determines a world of experience (see Figure 2).

Now if truth is relative and cannot be formed entirely by your will, then one traditional epistemological goal becomes impossible: evidence cannot be expected to produce agreement among different inquirers. But the notion of invariant truth unites agreement with another goal, getting to the truth, and when the possibility of getting to agreement is eliminated, the possibility of getting to the truth, even

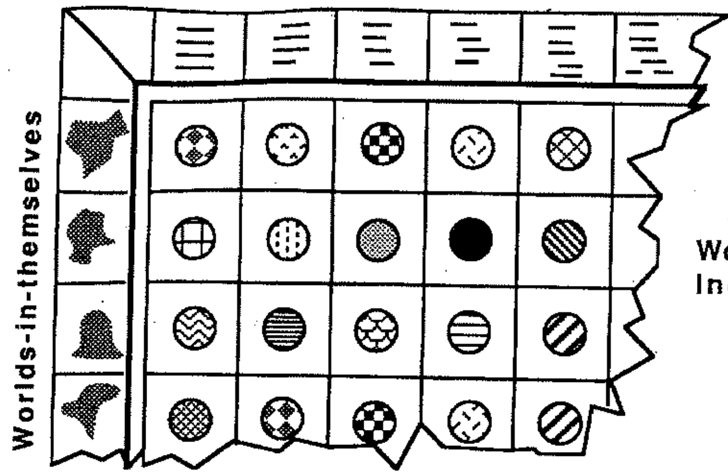



Fig. 2.

the relative truth, remains. A perfectly intelligible epistemological goal is to find the relative truth for you about some question. Since we do not want to presuppose anything about logic, let a question be given simply by some finite string $S$ over some finite set of elements. Each pair consisting of a world in itself and a conceptual scheme determines a status for the string: it is meaningful and true, meaningful and false, or meaningless. As the inquirer changes conceptual schemes the status of $S$ changes.

A whole range of questions suddenly appears. We can think of a discovery problem as given by a set of possible worlds in themselves and a set of possible conceptual schemes, with each member of their Cartesian product determining a world of experience. Suppose the inquirer receives data from any world of experience, just as in non-relativist discovery problems, but when he changes conceptual schemes, the world of experience from which he thereafter receives data also changes, depending on which world in itself is the actual one. Can the inquirer know the truth value of $S$ in the limit? That might mean: can he find a conceptual scheme in which $S$ has a truth value and stay in that conceptual scheme forever and converge to the correct truth value for $S$? Or it might mean: can he reach a point after which his changes of conceptual schemes have no effect on the truth value of $S$ and after which his conjectures about that truth value are correct? Or it might mean: is there a point after which $S$ always has a truth value and the inquirer always guesses the correct truth value for $S$, even though (because the inquirer changes conceptual schemes) the truth value of $S$ may change?

It is easy to construct simple examples of relativistic learning problems in which none of these kinds of knowledge can be obtained. Moreover, these different senses of knowledge in the limit are strictly inequivalent; there are problems that are solvable in the last sense but not in the second, and problems solvable in the second sense but not in the first. One can show that restrictions on conceptual schemes restrict the capacity for knowledge in the limit in each of these three senses. For example, there are problems involving an infinity of possible conceptual schemes that cannot be solved by any learner who is limited to a finite number of alternative conceptual schemes.

For the case in which the number of alternative conceptual schemes is finite, Kelly and I characterized the relativist discovery problems that are solvable in each of these three senses of "knowledge in the limit." For each conception of convergence there is a universal learner that will solve any problem solvable by any learner. In order to guarantee success, some fairly intricate strategies must be followed in deciding when to gather further evidence using a particular conceptual scheme and when to change conceptual schemes. If you believe yourself to be in a relativist system and your goal is to get to the relative truth for you, then the features of such strategies are epistemological norms.

Relativists might complain that they don't know which relativist system they are in, so they can't apply the norms, and a norm that cannot be applied is no norm at all. Can they learn which relativistic system they are in? Perhaps they think that which relativistic system one is in is relative to his conceptual scheme. Can one then learn the relative truth value of strings interpreted as claims about which relativistic system one is in? It would seem so in some cases if one follows the norms. But to follow the norms one must know which meta-relativistic system one is in. We can continue this way forever, just as with the Tarski language hierarchies. Unless a relativist thinks he can get out of the game, there is an epistemic norm for him.

These results only begin to touch the interesting questions about the epistemology of relative truth. Consider that much else could be relative to the inquirer's conceptual scheme, including the very history of the inquirer's conjectures. Consider the troubles that can result for those who attempt to learn theories *AE* in a relativistic system, when the truth is a function of the theory one conjectures.

## Applications

One person's application is another person's theory. What potential applications are there of these epistemological ideas to other enterprises?

### THE HISTORY OF PHILOSOPHY

The epistemological ideas about discovery that emerged from logic and the theory of computation are closely tied to history of philosophy, and they can be used to look back upon that history. The effect is to illuminate very different aspects than one finds in the histories of professional historians of philosophy. The convention, for example, is that Plato's Meno paradox is a paradox about reference. It is not. Bacon's *Novum Organum* is essentially a concept learning procedure, whose reliability can be described and compared with contemporary procedures. Kant's antimonies of reason are for the most part valid arguments about what cannot be known in the limit.

## PHILOSOPHY OF SCIENCE

What remains of general methodological discussions in philosophy of science consists largely either of arguments over "rational" relations between theory and evidence or historicist recomendations for assessing scientific traditions and research programs. If the principal point of inquiry is to get to the truth, or to get to certain kinds of truths, then these discussions typically establish nothing about the connections between the methodological notions that are advocated and the goal of inquiry. Considerations of when knowledge is and is not possible in the limit, and by which inferential strategies, keep the connection. Consider just a few examples.

Philosophers of science dispute when evidence is "relevant" to a hypothesis. There are probabilistic accounts that follow a subjectivist framework and treat evidence as relevant for someone if it changes his degree of belief in the hypothesis; there are logical accounts, such as hypothetico-deductivism and my own "bootstrap" account of evidential relevance. Each of these accounts looks like so much logical or probabilistic sociology, and the disputes among them often look like equivocations. Consider whether a class of possible evidence sentences is "relevant." If the goal is knowledge in the limit, and someone is following a particular strategy, a particular rule for conjecturing, then evidence in the class can be relevant for him for a particular discovery problem provided that his limiting behavior would be different if evidence from that class were deleted from each possible data sequence. In a more robust sense, a class of evidence is relevant to a discovery problem provided that the problem can be solved when that class of evidence is included in the data sequences, but when the problem is altered by removing evidence of that class for each sequence, knowledge in the limit can no longer be obtained. These features of evidential relevance turn out to be purely logical matters.

Methodologists dispute whether theories should always be consistent with the data and with background knowledge; whether the process of theory formation and alteration should be conservative and not make changes unless the current theory is contradicted by the evidence; whether theories should be simple in one or another sense. Each of these methodological principles will entail a cost for computationally bounded learners: there will be knowledge that can be obtained in the limit but not by any learner who abides by the methodological restrictions. Just where the costs lie remains to be investigated.

The extant results about knowledge in the limit connect directly with the concerns in philosophy of science that originally motivated Putnam's investigations. Putnam, recall, proved that for any "Carnapian" confirmation function for a sufficiently rich language there is a possible true sentence that never receives degree of confirmation as large as 1/2, no matter how much positive evidence of the hypothesis is presented. We can now see the same sort of thing much more generally. Suppose a probabilistic learner who changes probability distribution by conditionalizing on the evidence (or by any other means) converges to probability greater than 1/2 for a sentence $S$ if and only if that sentence is true. Then an obvious corollary of Kelly's characterization is that the evidence sequences satisfying $S$ must be $\Sigma 2$ in the appropriate hierarchy. For example, if the evidence is singular and the set of structures consists of all countable structures for the language, then the sentence must be logically equivalent to a sentence with a series of existential quantifiers followed by a series of universal quantifiers. So it is easy to give sentences and collections of possible structures such that no probabilistic learner can converge to probability greater than 1/2 in just the structures in which the sentence is true.

## ARTIFICIAL INTELLIGENCE

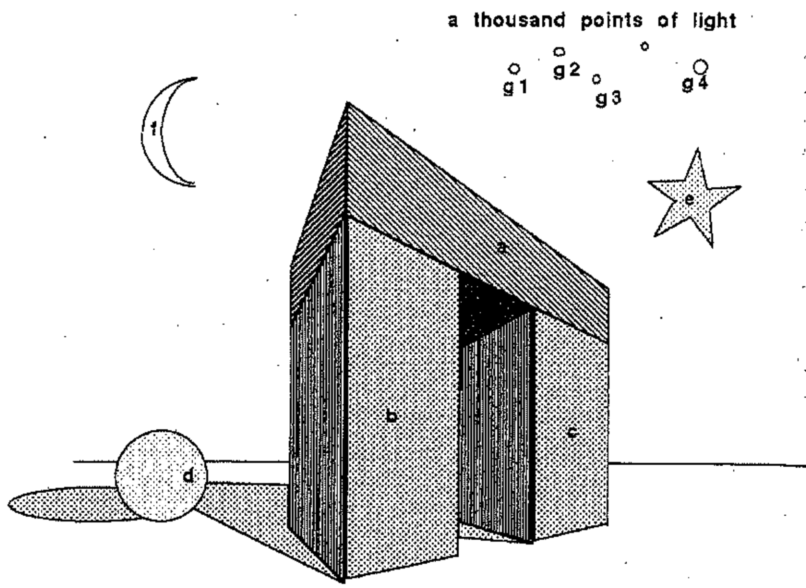
Thinking about limiting knowledge can sometimes be useful in understanding what a machine learning program does and doesn't do. One example will suffice.

Patrick Winston developed a well-known automated system for learning relational concepts from examples. The program will, for example, learn the concept of an arch from examples of facts about systems of blocks that are and are not arches (see Figure 3). In terms of non-logical predicates "$x$ is a block," "$x$ supports $y$," "$x$ touches $y$," "$z$ is a part of $u$," we could define *arch* by

> $\forall u \exists x \exists y \exists z \forall w [Arch(u) \leftrightarrow x$ is a part of $u$ and $y$ is a part of $u$ and $z$ is a part of $u$ and $x$ supports $z$ and $y$ supports $z$ and $x$ does not touch $y$ and if $w$ is a part of $u$ then $w$ is identical with $x$ or $w$ is identical with $y$ or $w$ is identical with $z$].

Consider whether any system could know in the limit whether or not this formula is true in a structure from data consisting of singular facts. Since the sentence is not $\Sigma 2$, we know that is impossible. How then does Winston's program manage to learn the concept? The answer is that the data the program is given is not confined to singular facts, but includes universal data. The program is told, for example, that a certain list of parts is *all* of the parts of an object. The hypothesis is $\Pi 1$ relative to universally quantified data.

The enterprises of "circumscription," "closed world assumptions" and so forth that has occupied so much effort in artificial intelligence appear to be simply a variety of methods for restricting the connection between data and hypotheses so that finite singular data will tacitly contain universal information. There is nothing




Arch(k)
Part-of(b,k)
Part-of(c,k)
Part-of(a,k)

Brick(a)
Brick(b)
Brick(c)
Vertical(b)
Vertical(c)
Horizontal(a)

Ontop(a,b)
Ontop(a,c)
Notouch(b,c)

"Closed world assumption"
(x)[if not (x=a or x=b or x=c) then not Part-of(x,k)]

Fig. 3.

objectionable in giving a machine (or a person) universally quantified data or data otherwise quantified, and one may for reasons of application be interested in finding one or another set of axioms that permit such information to be given indirectly through apparently singular data. But there is no reason to obscure the very simple epistemological structure at stake.

COGNITIVE SCIENCE

Mathematical cognitive psychology contains a number of "impossibility" theorems that assert the indistinguishability of certain hypotheses from evidence of certain kinds. Features of short term memory phenomena, for example, can provably be accounted for either by serial or by parallel processes. The literature on response times contains a number of such results.[9] Results of this sort are valuable in sorting out which of our allegiances are "working hypotheses" or "metaphysical background" for which we cannot hope to get empirical evidence of certain kind. They teach us that we must either be tolerant even as we pursue our conviction, or else we must look to other forms of evidence to establish our case.

One of the first applications of limiting analyses was of this sort. Gold considered a "black box" containing an unknown Turing machine. You can put an input into the box and you will get an output. You can repeat the process with different inputs, forever. Suppose after each trial you attempt to conjecture the future behavior of the machine. Is it possible to be right in the limit? Is there a strategy for conjecturing such that there will be some time after which the conjectures about the future behavior of the machine are correct. Gold proved there is not. If we have the computational power of Turing machines, then behavioral evidence cannot reliably predict behavior even in the limit. If, however, it is known that the black box can contain only some (unknown) finite automaton, then its behavior can be predicted.

A consequence of Kelly's characterization is a reflection that is almost intuitively obvious but so far as I know otherwise unremarked: It is impossible to determine from input output behavior whether or not a system is computationally bounded at all. That is, from data consisting of initial segments of the graph of an unknown function, one cannot reliably determine in the limit whether or not the function is computable.

## COGNITIVE NEUROPSYCHOLOGY

Cognitive neuropsychology aims to discover something about the functional architecture of human cognition principally from data about normal human capacities and abnormal incapacities. Schematically, the theories neuropsychologists produce are directed graphs with input vertices and output vertices. A capacity is a list of inputs and an output such that there is a path from each of the inputs to the output. Different capacities can overlap in their set of inputs, and different capacities can have the same output. The internal vertices of a hypothetical graph represent "functional modules" where cognitive processing is supposed to take place.

There are currently hot debates among neuropsychologists over the structure of inference and the relevance of evidence in neuropsychology. Some argue that the structure of testing is hypothetico-deductive, some that it is a matter of bootstrapping. Some argue that studies of statistical relations of incapacities in groups of subjects are relevant data, and some argue that they are not. Some argue that dissociations – the occurrence of an incapacity and a capacity together in an abnormal subject – are the most important data, others that double dissociations – the occurrence of an incapacity and a capacity in one subject and the reverse in another subject – are the crucial evidence. Some argue that associations – the fact that certain incapacities or capacities always occur together – are just as important as dissociations.

There is a natural structure in these issues that might usefully be clarified by thinking through the issues in terms of what can be known in the limit. The neuropsychologists' problems are about knowledge in the limit, rather than about Platonic knowledge, because they do not at any point know that the array of observed combinations of capacities and incapacities exhausts the possibilities. Misfortune might at any time present a new subject with a new combination. Depending on background assumptions, observed combinations can be used to exclude various architectures, and strategies that take advantage of our knowledge of learning in the limit may offer the possibility of increased reliability. At the very least, the learning theoretic framework should move the focus from arguments over methods of argument to the fundamental question of the reliability of inference and data acquisition strategies.

### ECONOMICS

One place in which a kind of relativism does obtain is the social sphere. What one does or says can have an effect on the truth value of what one claims. Consider only stock market prognosticators. Games have a similar feature, in which one player's expectations for an opponent's behavior depend on what the first player decides to do. Results about learning in the limit are a kind of a game in which the inquirer plays against a demon: the demon tries to deceive the learner in the limit, the learner tries not to be deceived. If there is a strategy for the learner such that the demon cannot succeed if the strategy is followed, we say the discovery problem is solvable; if there is a strategy the demon can follow such that no matter what strategy the learner follows he will be wrong in the limit, we say the discovery probem is unsolvable. In the relativist setting the relations between the inquirer and the demon are more nearly symmetrical. A completely symmetrical version of learning in the limit would be a setting for the investigation of infinite games, with and without computationally bounded players.

## Conclusion

There is a great deal more to be discovered about discovery, much of it undoubtedly not about knowledge in the limit. We should by all means seek to discover what can be known in the short run with sufficiently strong background knowledge and to understand how to measure the complexity of discovery and the interaction of probabilistic ideas with computation and complexity. But we ought not for a moment to take seriously the claim that there is no systematic, rigorous, informative theory of discovery. There is a very handsome, simple theory, and it has an excellent pedigree.

## Notes

[1] I am indebted to Kevin Kelly for several years of happy conversation from which the perspective and views of this paper grew, for comments on a draft of the paper, and for constructing some of the illustrations. A fellowship from the John Simon Guggenheim Memorial Foundation provided the liberty to write this paper. It was first presented in the Turing Colloquium, 1990.

[2] W. Newton-Smith, *The Rationality of Science*, Routledge and Kegan Paul, 1981, p. 125.

[3] For a more detailed discussion of the *Meno* see C. Glymour and K. Kelly, 'Thoroughly Modern Meno', in J. Earman, ed., *Pittsburgh Studies in Philosophy of Science*, Pittsburgh, PA: University of California Press, forthcoming.

[4] 'Trial and Error Predicates', *Journal of Symbolic Logic* **30** (1965), pp. 49–57. Gold's paper, 'Limiting Recursion' is in the same issue.

[5] See his essay on Carnap and 'Probability and Confirmation', Reprinted as chapters 17 and 18 of his collected papers.

[6] *Algorithmic Program Debugging*, M.I.T. Press, 1982.

[7] 'Convergence to the Truth and Nothing but the Truth', *Philosophy of Science*, 1989 and 'Theory Discovery from Quantified Data', *Journal of Philosophical Logic*, 1990.

[8] 'Paradigms of Truth Detection', *Journal of Philosophical Logic*, 1989.

[9] See R. D. Luce, *Response Times*, Oxford University Press, 1986.