Convergence to the Truth and Nothing but the Truth

Kevin T. Kelly; Clark Glymour

Philosophy of Science, Vol. 56, No. 2 (Jun., 1989), 185-220.

Stable URL:
http://links.jstor.org/sici?sici=0031-8248%28198906%2956%3A2%3C185%3ACTTTAN%3E2.0.CO%3B2-U

Philosophy of Science is currently published by The University of Chicago Press.

Philosophy of Science

June, 1989

CONVERGENCE TO THE TRUTH AND NOTHING BUT THE TRUTH*

KEVIN T. KELLY†

Department of Philosophy
Carnegie-Mellon University

CLARK GLYMOUR

Department of Philosophy
Carnegie-Mellon University
and
University of Pittsburgh

One construal of convergent realism is that for each clear question, scientific inquiry eventually answers it. In this paper we adapt the techniques of formal learning theory to determine in a precise manner the circumstances under which this ideal is achievable. In particular, we define two criteria of convergence to the truth on the basis of evidence. The first, which we call *EA* convergence, demands that the theorist converge to the complete truth "all at once". The second, which we call *AE* convergence, demands only that for every sentence in the theorist's language, there is a time at which the theorist settles the status of the sentence. The relative difficulties of these criteria are compared for effective and ineffective agents. We then examine in detail how the enrichment of an agent's hypothesis language makes the task of converging to the truth more difficult. In particular, we parametrize first-order languages by predicate and function symbol arity, presence or absence of identity, and quantifier prefix complexity. For nearly each choice of values of these parameters, we determine the senses in which effective and ineffective agents can converge to the complete truth on an arbitrary structure for the language. Finally, we sketch directions in which our learning theoretic setting can be generalized or made more realistic.

*Received February 1987; revised August 1987.
†We thank Dan Osherson for helpful comments on a draft of this paper.

Philosophy of Science, 56 (1989) pp. 185–220.

1. Introduction: Convergent Realism and Formal Learning Theory

> [A]ll the followers of science are animated by a cheerful hope that the process of investigation, if only pushed far enough, will give one certain solution to each question to which they apply it. . . . This great hope is embodied in the conception of truth and reality. The opinion which is fated to be ultimately agreed to by all who investigate is what we mean by the truth, and the object represented in this opinion is the real.
>
> . . . .
>
> [It] is unphilosophical to suppose that, with regard to any given question, (which has any clear meaning), investigation would not bring forth a solution of it, if it were carried far enough. (Peirce 1965, pp. 268–269)

Peirce held that reality is the theory to which we are disposed to converge in belief. And his sense of convergence seems to be this: for each clear proposition, there is a time at which its belief status is permanently settled by an ongoing inquiry.

Popper has proposed that as science progresses the corpus of scientific beliefs increases in verisimilitude (Popper 1963). As time goes on, scientific belief includes more of what is true and less of what is false; moreover, scientific belief increasingly contains more true *generalizations* and fewer false generalizations. Although attempts to define a verisimilitude relation statically have all met with logical difficulties, the notion of a convergent process still makes sense. Ideally, each true generalization will eventually be accepted, and continue to be accepted, and each false generalization will eventually be rejected and continue to be rejected. Although one cannot say of any two stages that one has more or less verisimilitude than the other, the process converges to the truth, and nothing but the truth.

In this paper, we investigate the conditions under which the convergence ideal implicit in the views of Peirce and Popper can be realized. That is, we investigate the languages for which there is a mode of inquiry such that for each proposition expressible in the language there is a time at which the mode of inquiry correctly settles its truth on the basis of evidence about particulars. We classify languages by predicate arity, by function symbol arity, by presence or absence of identity, and by quantifier prefix complexity.

2. Concepts of Convergence. In the formal study of learning, a learner is viewed as a function from evidence to hypotheses. If the function is recursive, we say that the learner is *effective*. A learner may be viewed

as learning functions, languages, relational structures, or theories. For historical reasons, most effort so far has been devoted to the study of language learning.

In language learning problems, the evidence consists of finite sequences of strings from the language to be learned. Any infinite sequence, containing every string in the language, and containing no strings not in the language, is a *text* for the language. The hypotheses can be understood as the name, or index, of the recursively enumerable set of strings of the language. Following Gold (1967), a learner is said to *identify* a text t if after receiving some initial segment t_n of t, the learner conjectures some correct index for the language, and continues to make the same conjecture forever after. A learner identifies a language if it identifies every text for the language, and a learner identifies a collection of languages if it identifies every language in the collection.

Many variations on this arrangement have been studied (Osherson, Stob and Weinsten 1986; Angluin and Smith 1982). Rather than requiring that the learner converge to a single index for a language, for example, one can adopt the weaker requirement that the learner converge after finite time to some equivalence class of indices. Rather than requiring that the language to which the learner converges be exactly the target language, one can require that it approximate the target in some regard. Rather than requiring that identification occur on all texts, one can require only that the learner identify some distinguished collection of texts, and so forth.

3. *EA* and *AE* Convergence. The usual notion of successful learning examined in the language acquisition literature is this: a learner is successful just in case *there is* a stage of inquiry after which *every* possible string is correctly classified by the learner as well-formed or not well-formed. Or more generally, *there is* a stage of inquiry after which *every* question of some specified kind is correctly answered by every subsequent conjecture of the agent. We refer to any identification criterion with this feature as an *EA convergence* criterion, where *EA* corresponds to the order of the existential and universal quantifiers over stages of inquiry and questions, respectively.

The sort of identification that corresponds to the one Peirce and Popper had in mind is weaker than *EA* identification. In this weaker sense, a learner identifies a language provided that for *every* well-formed string in the target language, *there is* a stage after which every language conjectured by the learner contains that string, and for every string not in the target language, there is a stage after which every language conjectured by the learner excludes that string. Or more generally, *for every* question of a specified kind, *there is* a stage of inquiry after which each conjecture of the learner correctly answers this question. We refer to con-

vergence criteria of this sort as *AE convergence* criteria. Notice that *AE* convergence differs from *EA* convergence merely by swapping the respective quantifiers over stages of inquiry and questions.

One reason why *AE* convergence has not been examined in the language learning literature is that *AE* language learning problems are all trivial, in the sense that every *AE* language learning problem is effectively solvable so long as the languages are countable.[1] The required learner merely conjectures at each stage that the target language consists of exactly the strings the learner has seen so far.

But when theories, rather than languages, are the objects to be learned, *AE* convergence is not at all trivial. We show below that there are simple learning problems that cannot be solved in the *AE* sense by any learner, and that there are learning problems that cannot be solved in the *EA* sense by any learner, but which can be solved effectively in the *AE* sense. Intuitively, the difference is this: the membership of a string in a given language is a singular fact that can be entailed by finite evidence. But a universal hypothesis is not entailed by any finite, singular evidence, so simply conjecturing the evidence at each stage does not suffice to settle the status of such a universal hypothesis in finite time.

Many of our formal arrangements follow Osherson and Weinstein (1986). Consider an arbitrary first-order language L on non-logical vocabulary V. We refer to arbitrary recursive subsets of L as *languages*. In practice, the sets we call languages are specified by particularly simple restrictions on the standard conditions of well-formedness for first-order formulas.

Let R be a countable relational structure for the unrestricted first-order language L. An *environment* is an ω-sequence of basic formulas (atomic or negated atomic formulas). The set of basic formulas occurring in an environment e is denoted $rng(e)$. Let the notation $R, h \models s$ abbreviate the claim that structure R satisfies formula s under assignment function h from variables of the language of s to domain elements of R. Environment e is *for* structure R if and only if there is an assignment h of the countably infinite collection of variables of L onto the domain of S such that $rng(e) = \{B : B \text{ is a basic formula and } R, h \models B\}$. The set of all finite sequences of basic formulas is denoted by SEQ; the set of basic formulas occurring in $\sigma \in SEQ$ is denoted $rng(\sigma)$. $\&s$ denotes the conjunction of basic formulas occurring in s, $\exists \&s$ denotes the existential closure of $\&s$, and e_n denotes the initial segment of length n of environment e.

An *L-theory* is an arbitrary collection of sentences in L, where L is a (possibly restricted) first-order language. The L-theory of a structure R

[1]A non-trivial notion of convergence slightly more restrictive than *AE* convergence has been studied in the context of language acquisition by J. Feldman (Biermann and Feldman 1972, p. 34).

for L is just the set of all sentences of L that are true in R. We view a learning function, f, as a total function from *SEQ* to the class of all recursively axiomatizable L-theories.

Let R be any countable structure for L. Let e be any environment for R, and let f be any learning function.

DEFINITION 1. *f EA converges to the L-theory of R on e* iff there is an n such that for all $m > n$, for each sentence $s \in L$, $f(e_m) \models s$ iff $R \models s$.

DEFINITION 2. *f AE converges to the L-theory of R on e* iff for each sentence s in L there is an n such that for all $m > n$, $f(e_m) \models s$ iff $R \models s$.

DEFINITION 3. *f EA(AE) identifies the L-theory of R* iff fEA(AE) converges to the L-theory of R on every environment for R.

For each structure whose L-theory is recursively axiomatizable, it is trivial that *some* learner identifies the structure's L-theory: the learning function need only output the L-theory of that structure for any input. But it is not a trivial matter whether there exists a learner that can identify each structure in a given collection of structures. So in the balance of this paper we think of discovery problems as *collections* of structures rather than as single structures. To solve such a problem, a prospective learner must eventually succeed no matter which structure in this collection is the actual structure under study.

DEFINITION 4. *fEA(AE) identifies collection K of structures with respect to L* iff for each structure $R \in K$, fEA(AE) identifies the L-theory of R.

DEFINITION 5. A collection K of structures is *EA(AE) identifiable with respect to L* iff there is some learning function that EA(AE) identifies K with respect to L.[2]

[2]Osherson and Weinstein consider a notion of *EA* identifiability for structures rather than for theories. An Osherson-Weinstein learning function takes relational structures as its values, rather than theories. An Osherson-Weinstein learner converges on environment e to a structure R just in case for all but finitely many n, $f(e_n) = R$. Given an equivalence relation r on a countable set of countable structures of a language, an Osherson-Weinstein learner r-identifies a collection K of structures if and only if for every structure S in K, for each environment e for S, g converges on e to a structure r-equivalent to S.

The concept of *EA* theory identification with respect to a language L developed in definitions 1 through 7 above can be understood as a special case of Osherson and Weinstein's r-identification of structures. Take r to be the relation of elementary equivalence with respect to language L. That is, two countable structures are r-equivalent if and only if the same sentences of L are true in both. Suppose fEA identifies collection K with respect to L. Define f' so that for each $\sigma \in SEQ$, $f'(\sigma)$ = the deductive closure of $f(\sigma)$ restricted to L. Next, let c be some choice function that chooses a countable model for any given,

In the standard literature in formal learning theory, a set of languages is taken to be a learning problem, and the objects of study are the classes of problems that can be solved in accordance with a certain criterion of identification. Since this notation makes the statement of theorems much more compact, we adopt an analogous system in which a problem is an ordered pair $\langle K,L\rangle$, where K is a set of structures and L is a language. We denote the class of problems solvable in the AE sense by the symbol $[AE]$, and the class of all problems solvable in the EA sense by $[EA]$. For each criterion of identification C, we define

DEFINITION 6. Problem $\langle K,L\rangle \in [C]$ iff K is C-identifiable with respect to L.

Clearly, $[EA]$ is a subset of $[AE]$. The following is another obvious relation between the two criteria.

PROPOSITION 1. fAE identifies collection K of structures with respect to L if and only if for each $s \in L$, fEA identifies K with respect to the singleton language $\{s\}$.

Proof. $\rightarrow$ Let fAE identify K with respect to L. Let $s \in L$, let $S \in K$, and let e be for S. Then for each $s \in L$ there is an n such that for each $m > n$, $f(e_n) \models s$ iff $D \models s$. Hence, fEA identifies K with respect to $\{s\}$.

$\leftarrow$ Suppose that for each $s \in L$, f can EA identify K with respect to each singleton language $\{s\}$. Then for each $S \in K$, environment e for S, and $s \in L$, there is an n such that for all $n' > n$, $f(e_{n'}) \models s$ iff $S \models s$. Hence fAE identifies K with respect to L.

It follows that if collection K is AE identifiable with respect to L, then for each $s \in L$, K is EA identifiable with respect to the language $\{s\}$.

The following observation will also be useful in what follows:

PROPOSITION 2. $\langle K,\{s\}\rangle \in [EA]$ iff $\langle K,\{-s\}\rangle \in [EA]$

Proof. Let $f_s EA$ identify K with respect to $\{s\}$. Define for all σ in SEQ, $f_{-s}(\sigma) = \{-s\}$ if and only if $f_s(\sigma)$ does not entail s.

---

consistent L-theory. Then the composition of c with f' identifies K in Osherson and Weinstein's sense. Conversely, let g identify K in Osherson and Weinstein's sense. Define f so that for each $\sigma \in SEQ$, $f(\sigma)$ is the complete L-theory of $g(\sigma)$. Clearly, fEA identifies K with respect to L.

Hence, the results of Osherson and Weinstein concerning structure identification may be applied to questions of EA identifiability. One of these results provides a general characterization of EA identifiability with respect to a language (Osherson and Weinstein 1986).

4. Effective Learning. So far, our learners have been assumed to be arbitrary functions from *SEQ* to recursively axiomatizable *L*-theories. But some kinds of functions may be of more immediate interest than others. For example, methodologists might restrict their attention to *computable* learners, since an inductive method ought to consist of a set of clear principles that can be followed step by step by finite beings devoid of arcane, magical abilities. Computability may ultimately draw the boundary on animal and machine abilities too narrowly, but most cognitive psychologists are betting otherwise.

To talk about effective learners, we must alter our basic notion of a learner, for theories are infinite objects and a finite device can output only finite objects in finite time. But of course all recursively axiomatizable theories are recursively enumerable, and recursively enumerable sets can be finitely described in a computationally useful way by means of procedures. Think of a theory as the set of all Gödel numbers of its members. A set is recursively enumerable if and only if it is the domain of some partial recursive function. Let W_i denote the domain of ϕ_i, which is in turn the partial recursive function computed by the Turing program which is Gödel numbered i. We can think of W_i as specifying a theory, namely, the set of all sentences whose Gödel numbers are W_i.

Let an effective learner be a computable function from *SEQ* to the natural numbers, where the theory conjectured is W_i, the domain of the partial function computed by the ith Turing machine. Now we can define effective theory identification as follows:

DEFINITION 7. A collection of structures is *effectively EA(AE)* identifiable with respect to L iff some effective learner *EA(AE)* identifies it with respect to L.

$\langle K,L\rangle \in$ [*EAe*]([*AEe*]) iff K is *effectively EA(AE)* identifiable with respect to L.

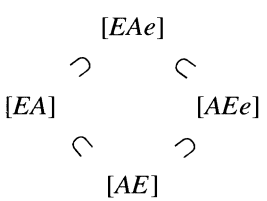
An immediate consequence of the original definition of *EA* identification is that only recursively enumerable theories are identifiable in that sense since the theory must be conjectured all at once and only recursively axiomatizable theories may be conjectured. But even effective learners can *AE*-identify a structure whose *L*-theory is not recursively axiomatizable. For let P be a monadic predicate, let s be a unary function symbol, and let 0 be an individual constant. Let L be the atomic sentences or negated atomic sentences over this vocabulary. Let R be a structure on the natural numbers that interprets P as a non-*RE* set of natural numbers, s as the successor function, and 0 as zero. The *L*-theory of R is not *RE* but a learner can *AE* converge to this theory by simply conjecturing the closed evidence sentences provided to it at each stage.

5. A Comparison of Identification Criteria. Recall that $[EA]$ is the set of all pairs $\langle L,K\rangle$ such that L is a restricted language, K is a set of structures for L, and there is a learner that EA-converges to the complete L-theory of each $R \in K$. $[AE]$ is the corresponding class that results when EA-convergence (that is, "all at once" convergence) is replaced with AE-convergence (that is, "piecemeal" convergence). $[EAe]$ and $[AEe]$ are just like $[EA]$ and $[AE]$, respectively, except the learners involved must all be effective.

We know that $[EA]$ is a subset of $[AE]$. It should be equally clear that $[EAe]$ is a subset of $[AEe]$, that $[AEe]$ is a subset of $[AE]$ and that $[EAe]$ is a subset of $[EA]$.

The next question is whether these inclusions are proper. Theorem 1 shows that $[EA]$ is not included in $[AEe]$, which implies that $[EA]$ is not included in $[EAe]$ and that $[AE]$ is not included in $[AEe]$.

The only question remaining is whether $[AEe]$ is a subset of $[EA]$. This question is answered negatively later in the paper by Propositions 3 and 9. Hence, we have the following complete picture of the inclusion structure of our four identification criteria:



Now we proceed to prove that $[EA]$ is not a subset of $[AEe]$.

THEOREM 1. There is a set of structures K and a restricted language L such that $\langle K,L\rangle \in [EA] - [AEe]$.

In the proof of Theorem 1 we construct a learning problem that "gives itself away" to a learner that can ineffectively decode its hints. But if there were an effective learner that could solve the problem, we could convert this learner into a limiting decision procedure for a set T that can be shown independently to have no such procedure. The trick is to define a procedure $foo(x)$ that produces complete evidence for a structure in which some sentence s is true if $x \in T$ and that produces complete evidence for a structure in which $-s$ is true if x is not in T. Then M's answers can be turned into a limiting decision procedure for T by dovetailing the search for proofs of s or $-s$ from the conjectures of M with the simulation of M on increasing initial segments of $foo(x)$.

Proof.

DEFINITION. $T = \{x : \phi_x \text{ is total}\}$

LEMMA a. T is Π_2-complete.[3]

Proof. First, T is clearly in Π_2, for x is in T just in case *for each* w *there* is a y such that $\phi_x(w)$ halts in y steps, and the matrix of the definition is a recursive relation. Note that the set $B = \{x : W_x$ is infinite$\}$ is Π_2 complete (Rogers 1987, p. 326). The following sort of reduction is also taken from Rogers (p. 326). Define a total recursive g such that for each z,

$$\phi_{g(z)}(x) = \begin{cases} 0 \text{ if } W_z \text{ has at least } x \text{ elements} \\ \text{divergent otherwise.} \end{cases}$$

That is, on input x the program $g(z)$ waits for the first x distinct integers to come out of the z enumeration of W_z and outputs 0 as soon as they are found. If they are never found, $g(z)$ will seek them forever, and hence will diverge, as is required. Then $g(z) \in T$ iff $z \in B$. Hence $T \geq_m B$. So T is Π_2 complete.

COROLLARY a. T is not limiting recursive.[4]

Proof. In Gold (1965) it is shown that the limiting recursive sets are just the Δ_2 sets. But T cannot be in Δ_2, since T is Π_2 complete and hence is not in Σ_2.

LANGUAGE. $L = \{\exists x \forall y Rxy, -\exists x \forall y Rxy\} \cup \{P[k] : k \in N\}$, where $P[k]$ is the usual sentence expressing that there are exactly n things with property P.

Let procedure *foo* be defined as in Figure 1. The evidence can include just three kinds of atoms. First, there are atoms of the form $P(x)$. Second, there are atoms of the form $R(x,y)$. Third, there are atoms of the form $x = y$, where x and y range over arbitrary variables in the language. It is apparent that for each selection of variables, each of the resulting atoms is either asserted on the output tape or denied on the output tape but never both. Since all identities among

[3]The following definitions are all from Rogers (1987). A quantifier prefix is said to be Π_n if and only if it begins with a universal quantifier and involves at most $n - 1$ alternations between blocks of universal quantifiers and blocks of existential quantifiers. A set of natural numbers is said to be Π_n if and only if it is definable by means of a Π_n quantifier prefix over a recursive relation. A set of natural numbers is Π_n-complete if and only if each Π_n set is many-one reducible to it. A set S is many-one reducible to a set S' if and only if there is a total recursive f such that for each natural number n, $n \in S$ if and only if $f(n) \in S'$. When S is many-one reducible to S' we write $S \leq_m S'$. The analogous definitions work for Σ_n, except that the quantifier prefix begins with an existential quantifier rather than a universal one.

[4]Set S is *limiting recursive* iff there is a total recursive g such that

$$\lim_y g(x,y) = \begin{cases} 1 \text{ if } x \in S \\ 0 \text{ otherwise.} \end{cases}$$

```
foo(k):
begin
  print 'P(x1)', . . ., 'P(xk)';
  go to stage 1;
end


stage n:
begin
  for each i, j ≤ n such that i is not equal to j
    do print '−xi = xj';
  for each i ≤ n do print 'xi = xi';
  for each i between k and n, print '−Pxi';
  for each i, j between 1 and n
  begin
    simulate φk(i) for j steps;
    if φk(i) halts in j steps or fewer,
    then print '−R(xi,xj)' on the tape
    else print 'R(xi,xj)' on the tape.
  end
  go to stage n + 1
end.
```

Figure 1. Procedure *foo*

distinct variables are denied, it is easily seen that there is a structure on the natural numbers that satisfies all formulas in the environment under interpretation $i(x_j) = j$. That is, the environment output by *foo* is *for* some structure.

LEMMA b. environment $foo(k)$ is for an S that satisfies $\exists y \forall x(Ryx)$ iff $k \notin T$.

If $k \in T$ then there is no computation $\phi_k(i)$ that does not terminate. Hence, for each i there is an n (that is, the number of steps it takes for ϕ_i to terminate) such that $R(x_i,x_j)$ occurs in $foo(k)$ only if $j \leq n$. So for each i there is a j (that is, $n + 1$) such that '$-R(x_i,x_j)$' occurs in $foo(k)$. Therefore, no structure for $foo(k)$ satisfies $\exists x \forall y Rxy$.

If k is not in T then there is some computation $\phi_k(i)$ that does not terminate. Hence, for each j, $R(x_i,x_j)$ occurs in $foo(k)$. Hence, $foo(k)$ is for a structure that satisfies $\exists x \forall y Rxy$. Q.E.D.

DEFINITION. Prob $= \{S : (Ex \in N)[foo(x) \text{ is for } S]\}$.

LEMMA c. Prob is *EA* solvable with respect to L.

Define learner M as follows. On a given finite segment σ of an environment e, M counts the distinct i such that $P(x_i)$ occurs in σ. Suppose this number is n. Then M decides ineffectively whether $n \in T$. If so, it conjectures the theory $\{P[n], -\exists x \forall y Pxy\}$. Otherwise, it conjectures the theory $\{P[n], \exists x \forall y Pxy\}$.

Intuitively, M assumes that the world under study satisfies each formula in environment $foo(n)$ with respect to some fixed interpretation, where n is the greatest number of distinct P things observed so far.

Since no structure in Prob has more than finitely many things with property P, there is a time k after which M is correct about the structure under study satisfying $foo(n)$ under some interpretation. Thereafter, by Lemma b, M produces exactly the true L-complete theory for the structure under study. Hence, M *EA* solves Prob with respect to L. Q.E.D.

LEMMA d. No effective learning function M *AE* solves Prob with respect to L.

Suppose there is such an M. We can convert M into an effective, limiting decision procedure for T, contrary to the first lemma. Consider the procedure *goo*, defined in Figure 2. To show that T is limiting recursive, we need only show (1), that *goo* is total recursive and (2), that

$$\lim_z goo(w,z) = \begin{cases} 1 \text{ if } w \in T \\ 0 \text{ otherwise.} \end{cases}$$

To see that *goo* is total recursive, note that *foo* is total recursive, and M is a partial recursive function that solves Prob. Since M solves Prob and $foo(w)$ is an environment for an element of Prob, $M(foo(w)_i)$ converges and the output entails either $\exists x \forall y Rxy$ or its negation for all but finitely many i. Hence there are always at least z k's such that $M(foo(w)_i)$ entails either $\exists x \forall y Rxy$ or its negation and the simple-minded dovetailing search of *goo* will find them.

To see that (2) is true, suppose first that $w \in T$. Then $foo(w)$ is for a structure S that does not satisfy '$\exists x \forall y Rxy$' by Lemma b. Since M *AE* identifies S w.r.t. L, M *AE* identifies environment $foo(w)$ for S w.r.t. L. Hence, all but finitely many conjectures of M on $foo(w)$ entail '$-\exists x \forall y Rxy$' and not '$\exists x \forall y Rxy$'. Hence, for all but finitely many z, $goo(w,z) = 1$.

Suppose w is not in T. Then $foo(w)$ is for a structure S that satisfies '$\exists x \forall y Rxy$' by Lemma b. By similar reasoning, there is a z after which $goo(w,z)$ always equals 0. Q.E.D.

6. Unrestricted Learning Problems. Recall that non-trivial inductive problems require a learner to identify the L theories of lots of different structures. Any set of structures for L constitutes such a problem, which may or may not be solvable. But for each language L, the maximally

```
goo(w,z):
begin
  set k: = 0;
  make TAPE 3 blank;
  until z answers are written on TAPE 3 do
  begin
    erase TAPE 1;
    let foo(w)_i denote the initial segment
      of foo(w) defined through stage i
      (c.f. the definition of foo);
    simulate the computations M(foo(w)_i)
      such that i < k for k steps each and
      write the outputs of these
      computations on TAPE 1 as they
      terminate;
    for each index m written on TAPE 1 do
    begin
      erase TAPE 2;
      use index m to list elements of
        W_m on TAPE 2
        for k computational steps;
      list all proofs of length k with
        premises written on TAPE 2
        to find a proof either of
        ∃x∀yPxy or of −∃x∀yPxy;
      if one of these proofs has
        conclusion −∃x∀yPxy
      then write 1 on TAPE 3;
      if one of these proofs has
        conclusion ∃x∀yPxy
      then write 0 on TAPE 3;
    end
    set k := k + 1
  end
  print the last entry on TAPE 3
end.
```

Figure 2. Procedure *goo*

difficult learning problem for which complete data can be presented is to identify each countable structure for L with respect to L. More precisely,

DEFINITION 8. For any restricted language L, $U(L)$ denotes the theory learning problem $\langle K(L),L\rangle$, where $K(L)$ is the class of all countable structures for L.

It is natural to expect that richer languages give rise to more difficult unrestricted learning problems. The more expressive a language, the finer the distinctions the learner must make to characterize his world in the language.

Clues to the expressive power of various syntactic features of a language can be obtained by reflection on the literature on the decision prob-

lem for first-order satisfiability. For example, the decision problem is sensitive to the arity of function symbols, the arity of predicates, and the complexity of quantifier prefixes (Dreben and Goldfarb 1979). It is therefore natural to look at *classes* of languages defined in terms of such restrictions in order to discover whether *each* language in a given class gives rise to an unsolvable theory learning problem.

DEFINITION 9. Let C be a class of first-order, prenex normal form languages. $C \in [[EA]]([[AE]]$, etc.) iff for each $L \in C$, $U(L) \in [EA]([AE]$, etc.).

Unless we mention otherwise, all languages considered in this paper will be assumed to have finite non-logical vocabularies and to include only sentences in prenex-normal form.

We consider sets of restricted languages characterized both by vocabulary and by quantifier prefix complexity. The properties of languages based on vocabulary restrictions are as follows:

$L \in F_n$ iff no function symbol of L is of arity greater than n.
$L \in P_n$ iff no predicate symbol of L is of arity greater than n.
$L \in I_1$ if L has the identity predicate and $L \in I_0$ otherwise.

There are many odd restrictions on the syntax of the matrix of a prenex formula that have an impact on expressiveness (Dreben and Goldfarb 1979) but we shall focus only on quantifier prefix form. We employ familiar notation due to Kleene to classify prenex formulas by quantifier prefix complexity. A formula is said to be Π_n just in case it is in prenex normal form and its prefix begins with a universal quantifier and involves at most $n - 1$ alternations between uninterrupted blocks of universal quantifiers and uninterrupted blocks of existential quantifiers. By convention, $\Sigma_0 = \Pi_0$ = the set of all quantifier free L-sentences. The definition of Σ_n is the same, except that the prefix must begin with an existential quantifier. We extend this notation to restricted languages in the following way:

$L \in \Pi_n$ iff each formula of L is a Π_n formula.
$L \in \Sigma_n$ iff each formula of L is a Π_n formula.

For brevity, we adopt the convention that for any two class names C, C', CC' denotes the intersection of C and C'. For example, $P_2F_0I_1$ denotes the set of all languages possibly involving identity, unary function symbols, and predicates of arity 2 or less. $P_2F_0I_1\Pi_1$ is the set of all such languages whose sentences have only purely universal quantifier prefixes. Notice that $P_2F_0I_1$ includes $P_2F_0I_1\Pi_n$, for each n, together with languages satisfying no finite quantifier alternation bound. Hence, a positive result for $C\Pi_n$ and $C\Sigma_n$, for all n, is weaker than a positive result for C.

The following table is a summary of our propositions concerning unrestricted learning problems.

POSITIVE RESULTS

Prop 3. $P_nF_0I_1\Pi_1 \in [[AEe]]$, for all n.

Prop 4. $P_nF_mI_1\Pi_1 \in [[AE]]$, for all n, m.

Prop 5. $P_nF_mI_1\Sigma_1 \in [[AEe]]$, for all n, m.

Prop 6. $P_1F_0I_0 \in [[EAe]]$

Prop 7. $P_1F_0I_1 \in [[AEe]]$

Prop 8. $P_1F_1I_0 \in [[AE]]$

NEGATIVE RESULTS

Props 9, 10. Neither $P_2F_0I_0\Pi_1$ nor $P_2F_0I_0\Sigma_1$ is in $[[EA]]$.

Prop 11. Neither $P_0F_0I_1\Pi_1$ nor $P_0F_0I_1\Sigma_1$ is in $[[EA]]$.

Cor 12, 12b. Neither $P_2F_0I_0\Sigma_2$ nor $P_2F_0I_0\Pi_2$ is in $[[AE]]$.

Cor 13. Neither $P_1F_2I_0\Pi_2$ nor $P_1F_2I_0\Sigma_2$ is in $[[AE]]$.

Cor 14. Neither $P_0F_1I_1\Pi_2$ nor $P_0F_1I_1\Sigma_2$ is in $[[AE]]$.

Prop 15. Neither $P_1F_1I_0\Pi_1$ nor $P_1F_1I_0\Sigma_1$ is in $[[EA]]$.

Notice that if $P_iF_jI_b\Pi_k$ or $P_iF_jI_b\Sigma_k$ fails to be in one of the identification classes under consideration, then no class of languages whose parameter values are respectively at least as great as i, j, b, and k is the class. Hence, the above results settle the membership questions for almost every class of languages characterized by these parameters. In fact, the only questions remaining are those that depend on the status of the following two open questions:

OPEN QUESTION 1. $P_nF_mI_1\Pi_1 \in [[AEe]]$, for all n, m?

OPEN QUESTION 2. $P_1F_1I_0 \in [[AEe]]$?

If both questions have affirmative answers, then we are done. But if either question has a negative answer, then there is likely a ragged line between positive and negative sub-cases. We expect that both questions have negative answers, and that settling the remaining sub-cases will be a substantial project.

Despite the open questions, the results listed above still provide a fairly comprehensive picture of the possibility of convergence to the truth over a variety of vocabularies and criteria of success. The picture is of particular interest in showing that naive methods that work for one hypothesis language fail dismally in another, slightly enriched language.

For example, Proposition 6 shows that learning hypotheses in the monadic predicate calculus is comparatively easy. This helps to explain the ubiquity of such problems in the artificial intelligence and cognitive psychology learning literature. Proposition 11 shows that adding identity to the language already thwarts *EA* learning, while Proposition 7 shows that effective *AE* learners can still handle the problem. Corollary 14 shows that adding one unary function symbol defeats both sorts of learners, even for purely universal or purely existential hypotheses. Corollary 13 shows that dropping identity and adding a binary function symbol also defeats both kinds of learners over purely universal and purely existential hypotheses.

The logical positivists observed long ago that existential hypotheses can be verified by finite data, universal hypotheses can be refuted on finite data, and hypotheses with mixed quantifier prefixes cannot be verified or refuted on finite data. That there is a close connection between these observations and *AE* learning is borne out by Propositions 3, 4, 5 and by Corollaries 12 and 12b. But notice that in the monadic predicate calculus, mixed quantification is no problem for *AE* learners (Propositions 6 and 7). The reason is that such sentences are "veri-futable" even though they are neither refutable nor verifiable.[5] It is also noteworthy that the ability to verify or to refute hypotheses is not sufficient for *EA* learnability (Propositions 9, 10 and 11).

It is tempting to see effective *AE* learnability as reflecting nothing more than the decidability of consistency between hypotheses and the evidence. But solvability of the decision problem for consistency does not imply solvability of the corresponding *AE* learning problem. For example, Proposition 13 shows that the unrestricted learning problem for the language consisting of the single sentence $\exists x \forall y Pxy$ is *AE* unsolvable. But the problem of deciding consistency between this sentence and any finite evidence (viewed as existentially closed) is solvable. For let $AE\Phi$ represent the hypothesis and $\exists \ldots \exists \& e$ be the existentially closed evidence. The question is whether $\forall\exists\Phi\&\exists \ldots \exists \& e$ is consistent. $\forall\exists\Phi\&\exists \ldots \exists \& e$ is equivalent to a sentence of form $\exists \ldots \exists\forall\exists(\Phi \& e)$. But it is known that deciding consistency over all function-free formulas on a given non-logical vocabulary whose prefixes are special cases of the form $\exists \ldots \exists\forall\exists \ldots \exists$ is a solvable problem (Dreben and Goldfarb 1979, p. 2) and $\exists \ldots \exists\forall\exists(\Phi \& e)$ is such a formula.

We do not know whether effective solvability of an unrestricted learning problem $U(L)$ implies solvability of the (consistency) decision problem for L. This question is closely tied to the Open Questions listed earlier. If the implication holds for ineffective agents, it is not because these

[5] "Verifutation" will be explained in detail in the proof of Proposition 6.

agents must employ a procedure to decide the consistency of their hypotheses with the data. And if the implication holds just for effective learners, it is not for the obvious reason that a successful, effective learner's conjectures must be consistent with the data. For example, the effective learner constructed in the proof of the next proposition does not always conjecture hypotheses consistent with the evidence, and may always fail to produce such a conjecture in some worlds.

On the other hand, if the implication fails, it cannot be shown false by the technique of Theorem 1. The technique in that proof is to exhibit a non-limiting recursive problem that would be limiting recursive if the *AE* learning problem in question were solvable by an effective agent. But first-order consistency is a Σ_1 problem, and hence *is* limiting recursive, so reducing it after the fashion of Theorem 1 would not show that the reducing *AE* learning problem is effectively unsolvable.

6.1 Positive Results

PROPOSITION 3. For all n, $P_nF_0I_1\Pi_1 \in [[AEe]]$

To prove this proposition we show that for each n and language $L \in P_nF_0I_1\Pi_1$ there is an effective procedure that *AE* identifies each structure in $U(L)$ with respect to L. The procedure is to increment a bin of universal hypotheses at each stage and to conjecture the set of all sentences in the bin that are not refuted by the available evidence at this stage. It is simple to show that for each universal sentence true in the structure to be identified, there is a point in the evidence after which the procedure always conjectures it, for at some point it enters the bin, and it cannot be refuted by the evidence thereafter, so it is always conjectured thereafter. The only trick is to show that for each false hypothesis there is a stage after which it is no longer entailed by any conjecture at any later stage. This is so because a purely universal sentence s false in a structure S is false in a finite substructure S' of S and hence some finite evidence characterizing S' will serve to refute any set H that entails s.

Proof. Let V be an arbitrary, enumerable, non-logical vocabulary without function symbols. Let L be the set of all purely universal closures of quantifier-free formulas over vocabulary V. Let τ be an effective enumeration of such formulas, and let τ_n be the nth formula in the enumeration.

Define the following procedure, where s is a sentence and $\sigma \in SEQ$, and where for any formula m and for any substitution function c from variables to variables, $m[c]$ denotes the formula that results from applying the substitution c to all occurrences in m of variables in the domain of c.

Procedure TEST(s,σ):
Form the matrix m of s.
For each c: var(s) $\rightarrow$ var(σ),
check by means of propositional logic
whether $m[c]$ is consistent with $rng(\sigma)$.
As soon as a c that makes these sentences
inconsistent is found, return "FAIL".
If no such c is found, return "PASS".

Procedure ID(σ):
set $n :=$ the length of σ; then conjecture
$\{s \in rng(\tau_n)$: TEST(s,σ) = "PASS"$\}$

Now, let S be a structure for L and let e be for S. Suppose $s \in L$ is true in S. For reductio, assume that s fails the TEST on e_n, for some n. Then there is a finite substitution c: var(s) $\rightarrow$ var(e_n) such that if m is the open matrix of s, then $m[c]$ is inconsistent with $\&e_n$. There is an h such that S, $h \models \&e_n$, since e is for S. But since $m[c]$ is inconsistent with $\&e_n$, S, h cannot satisfy $m[c]$, and hence S cannot satisfy s, which is a contradiction. Hence, TEST(s,e_n) = "PASS", for each n. Next notice that for each $s \in L$, there is a time t after which s is always tested and conjectured if it passes. After time t, s is entailed by each conjecture of ID, since s is in each such conjecture.

Now suppose s is false in S. Then there is some finite substructure S' of S in which s is false. Let h be a map such that $rng(e)$ = {basic formulas b: S,$h \models b$}, as is guaranteed by the fact that e is for S. Choose some finite set A of variables such that the h image of A is the domain of S'. There is one, since the domain of S' is finite. Let t be the least t' such that the set of all atoms of L in which only variables in A occur is a subset of $rng(e_{t'})$.

Let H be an arbitrary, finite subset of L such that $H \models s$. Since $H \models s$, there is some $s' \in H$ such that S' does not satisfy s'. Let m' be the matrix of s'. Since s' is false in S', there is an i: var(m') $\rightarrow$ domain of S' such that S', i do not satisfy m'. Choose c: var(m') $\rightarrow$ A such that h composed with c is identical to i, and such that the range of c is included in A. There is one, since the image of A under h is the whole domain of S', the range of i is a subset of the domain of S', and the domain of c is the same as the domain of i. Note that S', h satisfy $m'[c]$ iff S', $i \models m'$. Hence, S', h do not satisfy $m'[c]$. Let a be an atom occurring in $m'[c]$. a occurs in (e_t) iff S', $h \models a$, and $-a$ occurs in e_t otherwise, by the definitions of t and e. Since $\&e_t$ determines the truth value of each atom occurring in $m'[c]$, and

since $m'[c]$ is false under this valuation, $\&e_t$ and $m'[c]$ are inconsistent. Hence s' fails the TEST at each stage $t' > t$, and H would fail to be conjectured at any stage $t' > t$. And since t is fixed for all H such that $H \models s$, there is a t such that for each $t' > t$, H is not conjectured at t'. Q.E.D.

Nothing in the proof of Proposition 3 assumes that the non-logical vocabulary of L is finite. In fact, the constructions work so long as this vocabulary is countable.

In the proof of the previous proposition, we constructed for each false universal sentence s a bound n such that each set of purely universal sentences entailing s fails TEST after the nth evidence instance is read. This bound ensures that no false hypothesis is entailed infinitely often by the learner. This crucial bound arises because a purely universal sentence s false in a structure R is false in some substructure R' of R.

But when function symbols are added to the language, it is no longer true that when s is false in R, s is false in a finite substructure of R, for a structure with functions may have no proper substructures at all, let alone finite ones. Without such a bound, there are two difficulties to be avoided. First, consider a learner that conjectures the result of weeding down a BIN of sentences until the BIN is consistent with the current data. This learner may accidentally weed out some true sentence infinitely often to square its conjectures with the data, and hence may fail to AE converge to the complete truth. Next consider a learner that conjectures the result of weeding out of BIN all sentences with counterinstances in the data. This learner may produce infinitely many conjectures entailing a false hypothesis s, for the refuting instances may come too late to prevent combinations of late hypotheses entailing s from being added to BIN.

But there is a middle road that avoids both difficulties. On given evidence, the learner considers increasing initial segments of an enumeration of the hypothesis language. First it weeds out all hypotheses in this segment that have counterinstances in the data. Then it conjectures the greatest initial segment of the remaining sequence that is consistent with the current data. If a hypothesis is false, it is eventually refuted and no conjecture entailing it can be consistent with the evidence. If a hypothesis is true, all false hypotheses prior to it are eventually refuted, and it is included in the conjecture thereafter.

Open Question 1 asks whether there is an *effective* learner that duplicates this performance. The question remains open because the consistency test employed by the learner just described need not be effective.

PROPOSITION 4. For each n, k, $P_nF_kI_1\Pi_1 \in [[AE]]$

Proof. Let $n,k \in N$ and let $L \in P_nF_jI_1\Pi_1$. If $\sigma \in SEQ$, then let $\exists\&\sigma$ be the existential closure of the conjunction of the formulas occurring

in σ. Let τ be a fixed enumeration of L. Consider the following learner:

$f(\sigma)$:
begin
  set n:= length(σ);
  set c:= the result of deleting sentences from τ_n that
    have counterinstances in σ;
  conjecture the greatest initial segment of c consistent with
    $\exists\&\sigma$
end.

Let R be a countable structure for L and let e be an environment for R. Let s occur in position k of τ. Suppose s is true in R. There are at most finitely many false sentences preceding s in τ. Since these sentences are false and there are at most finitely many of them, there is a stage j after which each of them has a counterinstance in e_j. Let $n > \text{MAX}(j,k)$. At stage n, every sentence prior to position j in c is true. Hence, the greatest initial segment of c consistent with e_n must include s. So $f(e_n)$ includes s.

Suppose next that s is false in R. Then there is a j such that for all $j' > j$, a counterinstance to s occurs in $e_{j'}$. Suppose that $f(e_{j'}) \models s$. Then the greatest initial segment of c consistent with $e_{j'} \models s$. But this contradicts the fact that s is inconsistent with $e_{j'}$ in virtue of the supposed counterinstance. Q.E.D.

When the language is purely existential, things are easier. All the learner need do is to add existential formulas of the next greater size to the BIN if and only if they have instances in the evidence. Since true existential sentences eventually have such instances, and false ones cannot, this learner *AE* identifies any countable structure for such a language. Moreover, finding instances of existential formulas involves a simple search, so an effective learner can do the job. So we have the following proposition:

PROPOSITION 5. For all n, k, $P_nF_kI_1\Sigma_1 \in [[AEe]]$.

Next we show that the unrestricted learning problem for any monadic predicate language without function symbols or identity is in $[EAe]$. Notice that no restrictions on quantifier complexity are required for effective solvability.

PROPOSITION 6. $P_1F_0I_0 \in [[EAe]]$

The idea behind the proof of this proposition is that each sentence in the monadic predicate calculus can be viewed as a Boolean combination

of sentences with purely existential or with purely universal quantifier prefixes. Now it will not do to conjecture such a sentence until it is refuted or to wait to conjecture it until it is proved. But there is a simple combination of "verificationism" and "falsificationism" that the positivists may have overlooked.

Each complete theory in a monadic predicate language may be axiomatized by a conjunction whose conjuncts are quantifier free sentences, purely universal sentences and/or purely existential sentences. Following Hilbert (Hilbert and Bernays 1968), we say that such a sentence is in *primary normal form*. For example, let an axiom h be $A\&B\&C\&D$ where A and C stand for purely universal sentences and B and D stand for purely existential sentences. Let e be some singular evidence. Interpret the conjuncts of h in the following way: A purely existential conjunct is interpreted as T just in case e verifies it. A purely universal conjunct is interpreted as T just in case e does not refute it. The value of the whole conjunction is T just in case the value of each conjunct is. Thus, there is a distinct "verifutation" rule for each monadic predicate sentence expressed as a Boolean combination of purely universal and purely existentials. And there is a corresponding epistemic norm of "veri-falsificationism", namely, a conjecture is permissible only if its valuation with respect to the evidence is T. Our method is just to enumerate the (finite) set of primary normal form representatives of logical equivalence classes of complete theories and to conjecture on evidence e the first such axiom whose value with respect to the evidence is T. No false existential conjunct is ever verified, and hence an axiom involving such a conjunct is never conjectured. Each false universal conjunct is eventually refuted, so any axiom involving such a conjunct is conjectured only finitely often. If a hypothesis is true, then its universal conjuncts are always interpreted as true with respect to the evidence, and eventually the witnesses for its existential conjuncts arrive in the evidence. Once all false axioms prior to it in the order are refuted and the witnesses for its existential conjuncts arrive, it is conjectured forever after. The following proof is essentially an elaboration of the preceding observations.

Proof. Hilbert and Bernays (1968, pp. 145–146) show that an arbitrary sentence of monadic predicate logic can be rewritten as a Boolean combination of *primary* sentences, which are sentences of the following forms:

(1) $(\forall x)(B_1(x) \vee \ldots \vee B_n(x))$,

(2) $(\exists x)(B_1(x)\& \ldots \& B_n(x))$,

(3) $B_i(c_j)$

where $B_i(x)$ denotes either $P_i(x)$ or $-P_i(x)$, and where cj is a constant of L.

The set of all sentences of form (3) is finite. Call this set $V3$. Now consider the set $V1$ of all sentences of form (1) in which no negated predicate occurs negated more than once and no non-negated predicate occurs non-negated more than once. $V1$ is finite, (no member has more than $2n$ predicates occurring in it) and includes for each sentence of form (1) some sentence logically equivalent to it (since eliminating repeated disjuncts preserves equivalence). Let $V2$ be the result of applying the same restriction to sentences of form (2). $V2$ is likewise finite. Hence, $V = (V1 \cup V2 \cup V3)$ is finite.

Denote the atomic sentences of L by $A(L)$. Let $\Gamma = (V1 \cup A(L))$ and $s \in \Gamma$. Then we let neg(s) denote the element of $V2$ that results from negating s and driving the negation in if $s \in V1$, and the negation of s if $s \in A(L)$. Consider an arbitrary enumeration of Γ. A *pseudo-state* is a conjunction c such that the ith conjunct in c is either s or neg(s), where s is the ith element of Γ. There are just finitely many pseudo-states since Γ is finite. Also, every finite Boolean combination of sentences in V is equivalent to an irredundant, finite disjunction of pseudo-states. Sentences in this form may be said to be in *pseudo-state DNF*. Each sentence in L is equivalent to a disjunction of this form. Let S be an arbitrary relational structure for L and let T be the L-theory of S. Let $T' = \{s : s$ is in pseudo-state *DNF* and there is an $s' \in T$ such that s is logically equivalent to $s'\}$. T' is logically equivalent to T since each element of T is equivalent to some sentence in pseudo-state *DNF*. Let T'' be the result of choosing a true disjunct from each $s \in T'$. For each $s \in T'$ there is a pseudo-state disjunct of s that is true in S, since s is in pseudo-state *DNF* and $S \models s$. $S \models T''$ and $T'' \models T$ so T is logically equivalent to T'' and T'' is non-empty. Observe that no two distinct pseudo-states are mutually consistent, for if two pseudo-states differ, they differ by respectively asserting and denying at least one purely universal sentence. Since $S \models T''$, T'' is consistent and hence contains at most one pseudo-state. Since T'' is non-empty it contains at least one pseudo-state. Hence T'' is a singleton containing exactly one pseudo-state. Hence, (*) each complete L-theory is axiomatized by some consistent pseudo-state.

Let g be a consistent pseudo-state. Let $s \in L$. Let p be some primary normal sentence equivalent to s. g determines a Boolean valuation of s by entailing either r or $-r$, for each sentence $r \in V$ occurring in p. If the valuation so determined makes p true, then $g \models s$. If the valuation so determined makes p false, then $g \models -s$. And the valuation must either make p true or make p false since each

sentence in V or its negation is entailed by g. So (**) each consistent pseudo-state axiomatizes some complete L-theory. Hence, by (*) and (**), (***) there is a one to one correspondence between consistent, complete L-theories and logically equivalent, consistent pseudo-states.

Now we give an effective procedure that EA identifies the set of all structures for L. As before, if $\sigma \in SEQ$, then $\&\sigma$ denotes the conjunction of the formulas occurring in σ and $\exists\&\sigma$ denotes the existential closure of this conjunction.

Let PSEUDO be an enumeration of the consistent pseudo-states for L. Since the set is finite, it can be stored in a lookup table. (But since consistency for L is decidable, the list can be computed from the non-logical vocabulary of L as well.) Let $C_1\& \ldots \&C_n$ be an arbitrary element of PSEUDO. Now for an arbitrary $\sigma \in SEQ$, define the interpretation $i: V \to \{0,1\}$ for each $s \in V$ as follows:

$$i_\sigma(s) = \begin{cases} 1 \text{ if } s \in (V3 \cup V2) \text{ and } \exists\&\sigma \vDash s \text{ or} \\ \text{if } s \in V1 \text{ and } \exists\&\sigma \text{ is consistent with } s. \\ \\ 0 \text{ otherwise.} \end{cases}$$

Note that i_σ is effective, because consistency and entailment are decidable in the monadic predicate calculus. Now define learner f as follows:

$f(\sigma)$:
  conjecture the first element s of PSEUDO
  such that for each conjunct C occurring in s,
  $i_\sigma(C) = 1$.

Let R be an arbitrary, countable structure for L and let s be the first element of PSEUDO that is true in R. Let e be an environment for R. Since all s' prior to s in PSEUDO are false conjunctions, each has a false conjunct of type $V1$, $V2$ or $V3$. If the false conjunct C is in $V1$, then there is an n such that for all $m > n$, $\exists\&e_m$ is inconsistent with C and hence $i_{e_n}(s') = 0$ for each such m. If the false conjunct C is in $V2$ or $V3$ then there is no n such that $\exists\&e_n$ entails C, and hence for each n, $i_{e_n}(s') = 0$. Since there are finitely many competitors prior to s in PSEUDO, there is some finite n such that for all $m > n$, each false s' prior to s in PSEUDO has a conjunct C such that $i_{e_n}(C) = 0$. Call it n_a. Next, notice that for each n, $i_{e_n}(C) = 1$ for each conjunct $C \in V1$ of s, since s is true. Moreover, for each conjunct C of s in $(V2 \cup V3)$, there is some n such that for each $m > n$, $\exists\&e_m \vDash C$. Since there are finitely many conjuncts occurring in s, we have that there is an n such that for all $m > n$, $i_{e_m}(C) = 1$ for each conjunct occurring in s. Call it n_b. So for all $m > \text{MAX}(n_a, n_b)$,

$f(e_n)$ conjectures s. Hence f *EA* identifies every structure R for L. Q.E.D.

Intuitively, a system whose hypothesis language is the full first-order language on a vocabulary with only monadic predicates and identity can do no better than say exactly how many individuals satisfy each state description in the language. The method in the following proof exploits this observation.

PROPOSITION 7. $P_1F_0I_1 \in [[AEe]]$

Proof. Let σ be a finite segment of an L-environment. Say that σ *describes* term t iff σ consistently asserts or denies each monadic predicate of L of term t. For k monadic predicates, there are 2^k distinct ways to consistently assert or deny these predicates of a term. Enumerate these ways from 1 to 2^k. By "$D_i(t)$ occurs in σ" we mean that σ describes t in the ith of these ways.

Let $D[i]_\sigma$ denote the set of all terms occurring in σ that are described in way i by σ. In the following construction, the expression $(\exists!_n x)(\Phi(x))$ denotes the usual L-sentence stating that there are exactly n things with property Φ.

```
f(σ):
For each i from 1 to 2^k do
begin
   set D := D[i]_σ;
   set n := the cardinality of the least cardinality
      model of the negated identity atoms occurring in σ
   print '(∃!_n x)[D_i(x)]';
   for each constant c ∈ D[i]_σ do
      print 'D_i(c)'
end.
```

Clearly, f is effective, for n can be found by starting with a unit domain and adding objects until the negated identity atoms are satisfied, and all the rest of the searches are bounded.

Let structure S be a countable structure for L. Suppose S is finite. Let e be for S. Then there is an h such that S, h satisfies each formula of e. Moreover, there is a time n such that for each distinct domain element d of S, some term t denoting d with respect to h is described in e_n and for each constant c of L, c is described in e_n. Let T be the set of terms described in e_n. Then there is some time $n' > n$ after which each pair of terms in T denoting distinct domain elements of S with respect to h is dis-identified in $e_{n'}$. By the definition of f, the theory H conjectured by f does not change after stage n' and is true

of S by stage n'. But H determines the cardinality of the extension of each non-logical predicate, so any model of H is isomorphic to any other. Hence, the theory converged to is categorical and hence is complete. So f effectively EA converges to the complete theory of S with respect to L.

Now suppose that S is countably infinite. Then there is some D_k whose extension in S is infinite. Let DF be the set of all D_i for L whose extensions in S are finite. (DF may be empty, of course.) By reasoning similar to before, the theory AE converged to by f will correctly determine the cardinality of the extension of each state description with finite extension in S, since for each i such that $D_i \in DF$, the sentence $(\exists!_n x)[D_i(x)]$ is never rejected after some time. Now consider some D_j not in DF. As new descriptions and dis-identity statements feed into the evidence, f continually raises its estimate of the cardinality of D_j. Since lower lower bounds are entailed by higher lower bounds, each lower bound is eventually added and kept. But each upper bound is rejected when the estimate is raised, so each upper bound is eventually rejected. Hence the theory AE converged to entails each finite lower bound on the extension of D_j.

Clearly, the theory H thus converged to is true of S. The next question is whether it is complete. H has no finite model, since it entails every finite lower bound on the cardinality of the extension of D_j. But all countably infinite models of H are isomorphic, for these models must all assign extensions of the same finite cardinality to the same descriptions in DF and must all assign countably infinite extensions to the descriptions not in DF. Hence, by Vaught's test, H is complete. Q.E.D.

Finally, we consider the case of monadic predicates, unary functions, and no identity. The second Open Question results from the fact that the following method requires an entailment check that may not be effective. The method employed is an adaptation of the one involved in the proof of Proposition 4.

PROPOSITION 8. $P_1F_1I_0 \in [[AE]]$

Proof. Let $L \in P_1F_1I_0$. Then each sentence of L can be put into a logically equivalent primary normal form by the same procedure described in the proof of Proposition 6. Let L' be the primary normal fragment of L and let τ be an enumeration of L'. Let $\sigma \in SEQ$ and define the "evidential interpretation": i_σ just as in the proof of Proposition 6. Recall that for each purely universal or purely existential sentence s, function $i_\sigma(s)$ is either zero or one.

Next, define the evidential valuation $v_\sigma(s)$ of primary normal sen-

tence s with respect to evidence σ to be the truth value of s when its purely universal and purely existential components are interpreted according to i_σ.

LEMMA a. Let s be a sentence in primary normal form, let M be a structure for the vocabulary of s, let e be an environment for M. Then there is a j such that for all $j' > j$ $v_{e_{j'}}(s) = 1$ if $M \models s$ and there is a j such that for all $j' > j$ $v_{e_{j'}}(s) = 0$ otherwise.

Proof. A straightforward induction on the number of connectives occurring outside the scope of any quantifier in s. Q.E.D.

Our learner will examine two sets of primary normal form sentences defined as follows:

$$\text{PASS}(\sigma,m) = \{\tau_i : i \leq m \;\&\; v_\sigma(\tau_i) = 1\}$$

$$\text{FAIL}(\sigma,m) = \{\tau_i : i \leq m \;\&\; v_\sigma(\tau_i) = 0\}$$

Now we define a learning program g:

$g(\sigma)$:
begin
  set n := length(σ);
  set m := the greatest $k \leq n$ such that PASS(σ,k) does not entail
    any element of FAIL(σ,k);
  conjecture PASS(σ,m)
end.

Let M be a countable structure for L, let e be an environment for for M, let s be a sentence in L, and let τ_w be the first primary normal sentence in τ that is logically equivalent to s. Due to the lemma, there is an n such that for all $n' > n$ and for all s' prior to τ_w in τ, $v_{e_{n'}}(s') = 1$ if $M \models s'$ and $v_{e_{n'}}(s) = 0$ otherwise.

Since each element of PASS($e_{n'}$,w) = $\{\tau_i : i \leq w \;\&\; v_{e_{n'}}(\tau_i) = 1\}$ is true for all $n' > n$ and each element of FAIL($e_{n'}$,w) = $\{\tau_i : i \leq w \;\&\; v_{e_{n'}}(\tau_i) = 0\}$ is false for all $n' > n$, PASS($e_{n'}$,w) can entail no element of FAIL($e_{n'}$,w) for all $n' > n$. The learning program g does not even consider τ_w until it reads evidence e_w. But then the lemma and the definition of g yield

(*) for each $n' > \text{MAX}(n,w)$,
PASS($e_{n'}$,w) is included in the conjecture of g at stage n'
and no element of FAIL($e_{n'}$,w) is entailed by this conjecture.

Suppose that $M \models s$. So by the lemma, τ_w is in PASS($e_{n'}$,w) for each $n' > n$, since τ_w is logically equivalent to s. So by (*), s is entailed

by each conjecture of g after stage MAX(n,w).

Now suppose that s is false in M. Then for each stage $n' > n$, τ_w is in FAIL($e_{n'}$,w), by the lemma and the assumption that s is logically equivalent to τ_w. So by (*), τ_w and (hence s) is entailed by no conjecture of g after stage MAX(n,w). Q.E.D.

6.2. Negative Results. The negative results of this section complement the positive results of the previous section. For each class not shown to be in [[*EA*]] (or in [[*AE*]]) in the previous section, we show the class not to be in [[*EA*]] (or in [[*AE*]]) in this section. To put it another way, no *AE* learner constructed in the previous section can be "improved" to find the truth "all at once", rather than in "bits and pieces" for eternity. And no problem for which we did not find an *AE* learner has one. The Open Questions concern only the existence of effective *AE* learners where we have constructed ineffective ones.

First we show that even when the non-logical vocabulary consists of just a single binary relation R and even when disjunction is the only propositional connective admitted in the propositional matrix, the set of all structures for this vocabulary is not *EA* identifiable, even ineffectively, with respect to the purely universal closures of such formulas. The proof is an adaptation to our framework of Gold's familiar diagonal argument (Gold 1967). Even in the very restricted language just described, we can construct an infinite sequence of sentences such that each successor is entailed by its predecessor, but not conversely. Moreover, every finite segment of the complete evidence for a structure in which a predecessor is false but its successor is true can be extended to the complete evidence for a world in which the successor is itself false, but the successor's successor is true.

If we suppose for reductio that a learner can identify each structure for the language, then we can always "con" the learner into thinking that some sentence is false but its successor is true by presenting the appropriate evidence. At this point, we switch to a structure in which the successor is false but the successor's successor is true, and feed evidence from it until the learner is convinced, as it must eventually be, that the target is a structure in which the successor's successor is true. So we "con" the learner into changing its mind infinitely often on the evidence provided. Moreover, the total evidence provided is the environment for *some* structure for the hypothesis language, for each atom in the language occurs either negated or non-negated, but not both, in the resulting sequence, and any such sequence is the environment for some structure for the language. So the learner changes its mind infinitely often on an environment for some structure for the language, and hence fails to *EA* identify this structure, contrary to assumption.

PROPOSITION 9. $P_2F_0I_0\Pi_1$ is not in [[*EA*]].

Proof. Let the non-logical vocabulary be just the binary predicate 'P'. Let K be the set of all structures $\langle N;R\rangle$, where R is a relation on N^2 that interprets 'P'. Suppose f can EA identify K with respect to the purely universal closures of quantifier-free formulas on the non-logical vocabulary consisting of just the binary predicate 'P'. We construct an environment e for some structure S in K such that f changes its mind infinitely often on e, contrary to assumption. Let $s[n]$, for $n > 1$, denote the sentence

$$\forall x_1 \forall x_2 \ldots \forall x_n [Px_1x_2 \vee Px_2x_3 \vee \ldots \vee Px_{n-1}x_n].$$

Let $s[1]$ = any contradiction. For example, $s[2] = Ax_1Ax_2[Px_1x_2]$ and $s[3] = \forall x_1 \forall x_2 \forall x_3 [Px_1x_2 \vee Px_2x_3]$. Observe that for each n, $n' > n$, $s[n] \models s[n']$ does not entail $s[n]$. For consider $s[n]$, $s[n']$, $n' > n$. That $s[n] \models s[n']$ is obvious, since $s[n']$ is a case of weakening the matrix of $s[n]$. Now consider a structure S' whose domain is the natural numbers and whose relation R is just $N^2 - \{\langle i,j\rangle : Px_ix_j$ occurs in the matrix of $s[n]\}$. The sentence $s[n]$ is false in S' (consider the interpretation taking variable x_j to natural number j), but the sentence $s[n']$ is true in S' (at most, only $n - 1$ of the $n' > n - 1$ disjuncts of $s[n + 1]$ can be false in S' when $n > 1$).

Now define for each $n \in N$ and $\sigma \in SEQ$

$\mathrm{POS}[n] = \{Px_ix_j : (i = n \;\&\; j \leq n)$ or $(j = n \;\&\; i \leq n)\}$
$\mathrm{NEG}[n,\sigma] = (\mathrm{POS}[n] - \{Px_ix_n\}) \cup \{-Px_ix_n\}$, where i is the greatest k such that there is a y such that $-Pyx_k$ occurs in σ, if there is one, and is 1 otherwise.

Now we define e in stages such that at stage i, the finite sequence $e[i]$ is clamped onto the end of the previously defined sequence $e|i - 1$, where for each natural number n, $e|n$ denotes e restricted to its first n chunks, and where $e[i]$ is given as follows:

$$e[1] = \mathrm{POS}[1]$$

$$e[i] = \begin{cases} \mathrm{NEG}[i,(e|i-1)] \text{ if } f(e|i-1) \text{ entails } s[k] \\ \text{but not } s[k-1] \\ \text{and } k \text{ is the least } j > 1 \text{ such that} \\ \text{no counterinstance to } s[j] \text{ occurs in } e|i-1. \\ \\ \mathrm{POS}[i] \text{ otherwise.} \end{cases}$$

Notice that every pair $\langle x_i,x_j\rangle$ of variables of $i, j \leq k$ occurs in some atom or negated atom occurring in $e|k$. Hence, as we let k go to infinity, each atom or its negation occurs in e, but never both. So e is an environment for some structure $S \in K$.

LEMMA. Suppose fEA identifies the set K of all countable structures for L. Then for each n, there is an m such that

1. $f(e|m)$ entails $s[n]$
2. $f(e|m)$ does not entail $s[n-1]$
3. n = the least k such that no counterinstance to $s[k]$ occurs in $e|m$.

BASE. Suppose there is no m such that (1), (2) and (3) hold when $n = 2$. Then by the definition of $e[i]$, for each $i \in N$, $e|n = \{Px_ix_j : i,j \leq n\}$. Hence e is an environment for a structure S in which $s[2]$ is true. But unless f conjectures at some point in e a consistent theory that entails $s[2]$, f does not EA identify $S \in K$, contrary to assumption.

INDUCTION. Assume that for all $n' \leq n$ there is an m such that (1), (2) and (3) hold. Then by the definition of $e[i]$, $e[m+1] = \text{NEG}[m+1]$. By the definition of $\text{NEG}[m+1]$, a counterinstance to $s[n]$ now occurs in $e|m+1$, but no counterinstance to $s[n+1]$ does. And by the definition of $e[i]$, this situation is maintained until f conjectures a theory that entails $s[n+1]$ but not $s[n]$, for during this period, $e[i]$ is always $\text{POS}[i]$. Now suppose that there is no m' such that $f(e|m')$ entails $s[n+1]$ but not $s[n]$. Then $e[i] = \text{POS}[i]$, for all $i > m+1$. So in the limit, e is an environment only for structures S in which $s[n]$ is false but $s[n+1]$ is true, and hence f does not identify any of these $S \in K$, which is a contradiction. Hence, there is an m' such that (1), (2) and (3) hold for the case $n+1$.

Due to the lemma, the supposition that f identifies K implies that f does not identify some structure in K, which is a contradiction. Hence no f identifies K. Q.E.D.

The fact that purely existential sentences are equivalent to the negations of purely universal sentences may lead one to suspect that a very similar argument goes through in the existential case. This is true, so we merely highlight the changes in the following proof. Once again, we only need a single binary relation and the conjunction connective in the propositional matrix for the result to apply.

PROPOSITION 10. $P_2F_0I_0\Sigma_1$ is not in $[[EA]]$

Proof. let $s'[n] =$

$$\exists x_1 \ldots \exists x_n[-Rx_1x_2 \& -Rx_2x_3 \& \ldots \& -Rx_{n-1}x_n].$$

Let POS[i] and NEG[i,σ] be defined exactly as before, and define

$$e[1] = \text{POS}[1]$$

$$e[i] = \begin{cases} \text{NEG}[i,(e|n)] \text{ if } f(e|i-1) \text{ entails } s[k] \\ \text{but not } s[k+1] \text{ and } k \text{ is the least } j > 1 \text{ such that} \\ \text{some instance of } s[j] \text{ occurs in } e|i-1. \\ \\ \text{POS}[i] \text{ otherwise.} \end{cases}$$

The induction works just as before, except that now we fool f into making infinitely many distinct conjectures that turn out to be too weak, whereas before we fooled f into making infinitely many distinct conjectures that turned out to be too strong.

The general idea of a con game based on an infinite sequence of proper entailments can be applied to other vocabularies. For example, in the vocabulary consisting just of variables, identity, disjunction, and purely universal quantifier prefixes, one can construct a sequence of sentences such that the ith sentence says that there are at most i things. Clearly, the ith sentence entails the $i+1$th, but not conversely. For each point i, we can provide arbitrarily much evidence about a world in which there are at most i things, and it is clear that any finite segment of such evidence can be extended to evidence about a world in which there are at most $i+1$ things, $i+2$ things, and so forth. So we can always tease an allegedly general learner into believing that there are at most i things and then exhibit a new thing not identical to any he has seen before, for as many times as we please. The same construction fools a learner with respect to Σ_1 identity formulas. We can feed arbitrarily much evidence from a structure with n elements until the learner conjectures that there are at least n things but not that there are at least n' things for $n' > n$. If it doesn't eventually conjecture such a theory it misses the structure with just n things. Then we immediately switch to a structure with $n+1$ things, and so forth. The learner must change its mind infinitely often. These observations amount to a proof of the following proposition:

PROPOSITION 11.
$P_0F_0I_1\Pi_1$ is not in $[[EA]]$
$P_0F_0I_1\Sigma_1$ is not in $[[EA]]$

Now we have seen that $[EA]$ is a proper subset of $[AE]$. The proofs provide an explanation of this fact: the AE learner can afford to traverse the entire sequence of proper entailments constructed in the last two proofs, while the EA learner hasn't time to do so.

But AE identification soon meets its own match. One alternation between existential and universal quantifiers suffices to make the previous

unrestricted learning problem *AE* unsolvable, even by ineffective learners. Consider the case of Σ_2 prefixes, that is, prefixes of the form $\exists \ldots \exists\forall \ldots \forall$. The unrestricted problem for such prefixes is shown to be *AE* unsolvable by showing something much stronger: the set of all countably infinite structures for the vocabulary of the simple sentence $\{\exists x\forall y(Qxy)\}$ is not *EA* identifiable and hence is not *AE* identifiable, by Proposition 1.

The proof again relies on a con game. This time, we show the prospective learner lots of evidence about some x who seems to be related to everybody else. As soon as our mark takes the bait, we exhibit somebody not related to x and furthermore, we exhibit lots of y's who clearly fail to be related to everyone else. When our victim finally retracts his claim that somebody is related to everybody, we immediately display a new x' that for all the world seems related to everybody else, and continue in this manner, bedeviling the would-be learner for eternity.

PROPOSITION 12. The set of all countably infinite structures for the vocabulary of $L = \{\exists x\forall y(Qxy)\}$ is not *EA*-identifiable with respect to L.

COROLLARY 12. $P_2F_0I_0\Sigma_2$ is not in $[[AE]]$

Proof. Let D be some countably infinite set. Then let problem K be the class of structures $\{\langle D;R\rangle : R \text{ is a subset of } D^2\}$, where R interprets Q. Suppose for reductio that some learner f can identify K. We construct an environment e for a structure $S \in K$ on which f does not *AE* identify S, contrary to assumption. Let s denote the sentence '$\exists x\forall yQxy$'. Define e in stages such that at stage i, the chunk $e[i]$ is added to $e|i-1$, where for each natural number n, $e|n$ denotes e restricted to its first n chunks, where $e|0$ is defined to be the empty sequence and $e[i]$ is defined as follows:

$$e[i] = \begin{cases} \{-Q(x_i,x_k) : k \leq i\} \cup \{-Q(x_k,x_i : k \leq i\} \\ \text{if } f[e|i-1] \models s. \\ \\ \{Q(x_i,x_k) : k \leq i\} \cup \{Q(x_k,x_i : k \leq i\} \\ \text{otherwise.} \end{cases}$$

Order chunk $e[i]$ however you please, say lexically, before clamping it onto $e|i-1$ to form $e|n$. By continuing this process forever we arrive at the environment e.

Since fEA identifies R with respect to L, it follows that either (A) $f[e_n] \models s$ for all but finitely many n or (B) $f[e_n]$ does not entail s for all but finitely many n. But in case (A), any countably infinite structure for e must fail to satisfy s; for notice that no matter what f does, e is an environment for some countably infinite structure for

L. By the definition of e, e is only for structures in which at most finitely many domain elements are related to at most finitely many others, so no element can be related to infinitely many others, and hence no element is related to every other. Hence in case (A), f does not identify any countably infinite R for which e is an environment.

In case (B), any countably infinite structure for e satisfies s; for in this case, e says that all but finitely many domain elements are related to every other, so some domain element is related to every other a fortiori. And again, there are countably infinite structures for e, none of which is identifiable by f, which is contrary to the reductio hypothesis.

By Proposition 2 we have as an immediate corollary:

COROLLARY 12b. $P_2F_0I_0\Pi_2$ is not in $[[AE]]$

Constructions similar to the one in the proof of Proposition 12 yield the following results:

PROPOSITION 13. The set K of all countably infinite structures for the singleton language $L = \{\exists x \forall y P(f(x,y))\}$ is not AE identifiable with respect to L.

COROLLARY 13.
$P_1F_2I_0\Sigma_2$ is not in $[[AE]]$
$P_1F_2I_0\Pi_2$ is not in $[[AE]]$

Proof. Let '$P(f(x,y))$' encode $P(x,y)$ in the proof of Proposition 12. The second corollary follows from Propositions 2 and 13.

PROPOSITION 14. The set K of all countably infinite structures for the the singleton language $L = \{\forall x \exists y(f(y) = x)\}$ is not AE identifiable with respect to L.

COROLLARY 14.
$P_0F_1I_1\Pi_2$ is not in $[[AE]]$
$P_0F_1I_1\Sigma_2$ is not in $[[AE]]$

Proof. This proof is analogous to the one offered for Proposition 12. Let MAE identify K. Informally, we can begin to present the environment for a structure $R \in K$ with a function f that is onto some proper subset of the domain of R. Eventually, M must produce a conjecture that does not entail $\forall x \exists y f(y) = x$. At this point, we switch to a structure R' whose function f' is just like f wherever f was described in the evidence, but which is onto the domain. Eventually, M must produce a conjecture that entails $\forall x \exists y f(y) = x$. Then we switch to a structure R'' whose function f'' is just like f' wherever

f' was described in the data, but which is not onto the domain. Since only finitely many values of f can be described by each stage and the domain is infinite, this process can continue forever. Hence M does not AE converge to any theory on an environment for an element of K, which is a contradiction.

The first corollary is immediate. The second follows from Propositions 2 and 14.

The following result involves a construction similar to that employed in the proof of Proposition 9.

PROPOSITION 15. Neither $P_1F_1I_0\Pi_1$ nor $P_1F_1I_0\Sigma_1$ is in $[[EA]]$.

Proof. Let L be the Π_1 sentences over the non-logical vocabulary $\{P,f\}$. Clearly, $L \in P_1F_1I_0\Pi_1$. Consider the following sequence of sentences of L:

$$\forall x(Px),\ \forall x(P(f(x))),\ \forall x(P(ff(x))),\ \forall x(P(fff(x))),\ \ldots$$

In general, let $s[j]$ be the sentence $\forall x(P(f\ \ldots\ f(x)))$ in which f is composed j times with itself. If $i < j$ then $s[i]$ entails $s[j]$ but $s[j]$ does not entail $s[i]$.

Suppose g can identify each countable structure for L. Again, we recursively define a malicious environment that g cannot identify. Let τ be a complete enumeration of the terms of L. Now we define:

$$e|0 = P(\tau_0)$$

$$e|i = \begin{cases} (e|i-1) * \text{`}{-}P(t)\text{'} \\ \text{if } f(e|i-1) \text{ entails } s[k] \\ \text{but not } s[k+1], \text{ where} \\ k \text{ is the least } j \in N \text{ such} \\ \text{that (1) no counterinstance} \\ \text{to } s[j] \text{ occurs in } e|i-1, \\ \text{(2) } t \text{ is the first term in} \\ \tau \text{ such that } s[k] \text{ is a} \\ \text{variable renaming variant} \\ \text{of `}P(t)\text{', and (3) } t \text{ does not} \\ \text{occur in } e|i-1. \\ \\ (e|i-1) * \text{`}P(t)\text{' otherwise, where} \\ t \text{ is the first term in } \tau \\ \text{not occurring in } e|i-1. \end{cases}$$

Whatever g does, this environment is for a countable structure for L. But by an induction just like that in the proof of Proposition 9, g fails to EA converge to a theory on the environment so constructed,

which contradicts the hypothesis. Hence, f does not EA identify each countable structure for L. Q.E.D.

7. Theme and Variations. The results presented in this paper provide a systematic but idealized picture of how the expressive power of a learner's hypothesis language is related to the difficulty of the inductive problem the learner faces. Despite the many idealizations involved, our results apply immediately to learning studies in artificial intelligence and cognitive psychology, since the learning problems considered in these disciplines are typically phrased in some notational variant of a first-order language.

The general approach underlying this study can be extended to the investigation of more complicated inductive settings. Although the results will change when the framework is changed, our present results may be of use in focusing attention on questions and distinctions that might otherwise be missed. Moreover, examining a variety of settings, some of them idealized, can help to isolate the factors that make more realistic problems more or less difficult. In this way, a deeper understanding of realistic learning problems results from the study of more idealized ones.

To illustrate the flexibility of the learning-theoretic approach we sketch some obvious extensions of the settings studied in this paper. In general, we view a learning problem as involving at least the following elements:

1. hypothesis language L
2. evidence language L'
3. a set W of possible worlds
4. a relation of theory adequacy (that is, a relation in $W \times L$)
5. some sort of process generating evidence from worlds
6. a learner that somehow gets evidence about the world and conjectures L-theories.

From this point of view, the problems considered in this paper are of a very specific sort. In the balance of the paper we sketch some obvious variations.

7.1. The Evidence Presentation. The manner in which we present data to the learner in this paper is highly idealized. In our setting, which follows most work on language and function identification, the world *non-deterministically* produces a complete, true, sequence of quantifier-free evidence sentences.[6] Alternative settings are numerous.

[6]We say "non-deterministically", for the learner must expect any order of data, and there is no probability measure over the possible data orders for a given world. Where there is such a measure, we may say that the data generation process is stochastic, and if there is but one data presentation per world, the presentation process is deterministic. This distinction accords naturally with usage in the theory of automata.

For example, we could assume that the data presentation mechanism is stochastic. And we could also figure in a certain probability for false evidence sentences to appear in the presentation.[7] We could also examine the effect of more informative data. For example, what if the evidence language includes formulas with Π_1 prefixes? How about Π_2 prefixes? For each type of evidence, there will be new positive and negative results. Such problems can be viewed as idealizations of the problem facing the scientist who builds on the reliable theoretical work of his ancestors and the experimental work of his contemporaries to formulate a more comprehensive theory of the world.

Another variation would be to alter the setting to change the evidence protocol so that the learner asks questions that the world answers (perhaps with a certain probability of error). It is easy to see that the mere ability to ask questions cannot increase inductive scope, for given any question-asker, there is a passive learner that simulates it and feeds it the answer to its questions as they are passively encountered. But it is equally easy to see that an experimenter who puts questions to the world can succeed arbitrarily more quickly than a passive observer—who may have to wait centuries for a crucial bit of evidence.

The artificial intelligence literature even suggests a way to view learners making queries in a language with function symbols as finding physically necessary experimental laws (Carbonell 1987). The trick is to view function symbols modally, as denoting *abilities* of the experimenter. When the experimenter asks the oracle about the truth of some atom with a closed term, say '$P(f(g(c)))$', we view him as having performed the procedure $\lambda x f(g(x))$ on c. If 'f' corresponds to the experimenter's ability to paint and g corresponds to his ability to scrape, then when the experimenter queries '$P(f(g(c)))$' we interpret the situation as the experimenter actualizing a possible world in which c is scraped and painted.[8] Since the ability to query does not affect learnability, the results concerning the

[7]This sort of study is well under way in the case of Boolean concept learning (Kearns, Li, Pitt, and Valiant 1987).

[8]Technically, let M be a countable relational structure for a language L with only unary function symbols such that (1) each closed term of L denotes a distinct object in M and (2) each individual in M is denoted by some closed term of L. Then we can define possible worlds with respect to M as follows: M' is a possible world with respect to M iff (a) the domain of M' is the set of individuals in M denoted by some set of closed terms of L in which each constant occurs exactly once, (b) M' has no functions, and (c) the relations of M' are the relations of M restricted to the domain of M'. That is, a possible world is a sub-structure that can be gotten to by performing countably many finite experiments, where a finite experiment is a finite application of function symbols to a constant. Notice that the functions in M are not in possible worlds. This is proper, because functions represent experimental *abilities* connecting possible worlds rather than entities in possible worlds. So in this setting, function symbols may be thought of in much the way "box" is thought of in standard modal logic.

learnability of theories with function symbols can be interpreted as applying to the learnability of necessary experimental laws of this sort.

7.2. Possible Worlds and Background Knowledge. Much artificial intelligence work focuses on the issue of bringing background knowledge to bear on learning and other tasks. It is important to note that nothing about our framework assumes that the learner is ignorant. The structures in a learning problem may all be models of a given theory or they may even be models of some theory inexpressible in the agent's hypothesis language. To determine the theories with elementary classes that are solvable will require greater attention to model theory than was demanded for the results presented here.

There is, however, a sense in which our framework is not general enough to address the question of how to use background knowledge, but the required modification is easy and obvious. If a problem is just a set of structures, then there is no pressure on the methodologist to design general mechanisms to redirect the scope of a single method in light of a given background theory. We think it would be particularly interesting to look at inductive problems as classes of theories, so that a learner is given a theory in this class and is expected to identify each of its countable models. An examination of the complexity of such problems would yield theorems about questions some artificial intelligence practitioners consider incapable of formal analysis (Schank 1986).

7.3. Theory Adequacy. Insistence on convergence to true theories is a major idealization in the present framework. For example, a *verisimilar* theory often suffices in lieu of a true one, so it is natural to study the difficulty of converging in various senses to theories verisimilar to a given degree. There is already some very interesting work on learning concept descriptions to within a given degree of error (Kearns, Li, Pitt and Valiant 1987). It is an interesting question whether this sort of analysis can be generalized to apply to the problems examined in this paper. If it can be, then we expect much more optimistic results when even quite small errors are permitted in hypotheses.

7.4. Convergence. Many methodologists insist that to learn "in the limit" is to learn too late. We are not quite so strident, for what can be learned in a short time may be too little, and it is not clear that it is always better to learn too little in a short time than it is to learn a lot in a long time. All that is clear is that we shouldn't take more time to learn than we *have* to.

But regardless of one's opinions on this issue, it is important to know what sorts of learning problems can be solved in a relatively short time

to see exactly what the trade-off between scope and complexity is. For simple Boolean concept-learning tasks, such questions are already being answered from a learning-theoretic perspective (Kearns, Li, Pitt and Valiant 1987). Such results will not apply directly to the present framework, but it is once again an interesting question whether they have generalizations that do apply.

7.5. Summary. The general perspective of formal learning theory is a powerful guide in the examination of methodological questions. It provides answers where the traditional methodological literature does not even locate questions. Its sharp focus breathes new life into weary methodological truisms. And formal learning theory makes the logic of discovery look, at last, like logic.

REFERENCES

Angluin, D. and Smith, C. H. (1982), *A Survey of Inductive Inference Methods*. Technical Report 250, Yale University.

Biermann, A. W. and Feldman, J. A. (1972), "A Survey of Results in Grammatical Inference", in S. Watanabe (ed.), *Frontiers of Pattern Recognition*. New York: Academic Press, pp. 32–43.

Carbonell, J. and Gil, Y. (1987), "Learning by Experimentation", in M. B. Morgan (ed.), *Proceedings of the Fourth International Workshop on Machine Learning*. Los Altos, CA: Morgan Kaufmann, pp. 256–266.

Dreben, B. and Goldfarb, W. (1979), *The Decision Problem: Solvable Classes of Quantificational Formulas*. Reading, Mass.: Addison-Wesley.

Gold, E. M. (1965), "Limiting Recursion", *Journal of Symbolic Logic 30*: 28–48.

———. (1967), "Language Identification in the Limit", *Information and Control 10*: 447–474.

Hilbert, D. and Bernays, P. (1968), *Grundlagen der Mathematik*. Berlin: Springer-Verlag.

Kearns, M.; Li, M.; Pitt, L.; and Valiant, L. (1987), "Recent Results on Boolean Concept Learning", in M. B. Morgan (ed.), *Proceedings of the Fourth International Workshop on Machine Learning*. Los Altos, CA: Morgan Kaufmann, pp. 337–352.

Osherson, D.; Stob, M.; and Weinstein, S. (1986), *Systems that Learn*. Cambridge, Mass.: MIT Press.

Osherson, D., and Weinstein, S. (1986), "Identification in the Limit of First Order Structures", *Journal of Philosophical Logic 15*: 44–81.

———. (1989), "Identifiable Collections of Countable Structures", *Philosophy of Science 56*: 94–105.

Peirce, C. S. (1965), *The Collected Papers of Charles Sanders Peirce*, vol. 5. C. Hartshorne and P. Weiss, (eds.). Cambridge, Mass.: Belknap Press.

Popper, K. R. (1963), *Conjectures and Refutations*. New York: Harper and Row.

Rogers, H. (1987), *Theory of Recursive Functions and Effective Computability*. Cambridge, Mass.: MIT Press.

Schank, R.; Collins, G.; and Hunter, L. (1986), "Transcending Inductive Category Formation in Learning", *Brain and Behavioral Sciences 9*: 639–686.