(2)

AD-A247 046

IMENTATION PAGE

Form Approved
OMB No. 0704-0188

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
January, 1992	Final 1 July, 89 - 31 Dec., 91

4. TITLE AND SUBTITLE
On the Automated Discovery of Scientific Theories

5. FUNDING NUMBERS
C N00014-89-J-1725

6. AUTHOR(S)
Daniel Osherson and Scott Weinstein

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Pennsylvania

8. PERFORMING ORGANIZATION REPORT NUMBER
IRCS-91-46

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ONR, 800 N. Quincy St.,
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
To be published in: Susan Chipman and Alan Meyrowitz (eds.)
Machine Learning: Induction, Analogy, and Discovery, Kluwer Academic Publishers

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This paper summarizes recent research results on applications of computational learning theory to problems involving rich systems of knowledge representation, in particular, first-order logic and extensions thereof.

14. SUBJECT TERMS
Machine Inductive Inference, Computational Learning Theory

15. NUMBER OF PAGES
18

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

On the Automated Discovery of Scientific Theories*

Daniel Osherson
IDIAP

Scott Weinstein
University of Pennsylvania

Abstract

This paper summarizes recent research results on applications of computational learning theory to problems involving rich systems of knowledge representation, in particular, first-order logic and extensions thereof.

INTRODUCTION

Science is such a useful activity that people have become interested in automating it, at least in part. A great deal of fruitful effort has been devoted to this task (e.g., [9, 10, 30]) but the limitations of existing systems lead to reflection about the very character of empirical inquiry. What is a scientific theory, after all, and what makes one theory better or worse than another? How should inquiry proceed in order to maximize our chances of believing a true theory, and minimize the chance of believing a false one? And how much success can be expected of the scientific enterprise, especially when carried out with limited access to data? Consideration of such matters leads to a set of interlocking issues at the heart of contemporary epistemology, including questions about probability, simplicity, approximate truth, hypothetical entities,

*Research support was provided by the Office of Naval Research under contracts Nos. N00014-87-K-0401 and N00014-89-J-1725. Correspondence to D. Osherson, IDIAP, C.P. 609, CH-1920 Martigny, Switzerland, *e-mail:* osherson@idiap.ch

and rational belief. Already resistant to clarification and solution, these questions become even more difficult when scientists are conceived as resource-limited computational agents

Faced with such conceptual complexity, the natural strategy is to experiment with alternative, simplifying assumptions about scientific practice and attempt to derive general theorems about empirical inquiry within the simpler contexts so defined. It may then be hoped that comparison and analysis of the results obtained will lead to insights that bear on the practical problem of building artificial systems of empirical inquiry in science, industry, medicine, etc. While there is no guarantee that such a research strategy will succeed, we note that it is analogous to past endeavours whose impact on technology have been substantial (for example, the analysis of alternative models of computation).

Thus is born the discipline of Computational Learning Theory whose goal is to define and analyze increasingly realistic models of empirical inquiry.[1] Each such model is adapted to a particular *discovery problem,* by which we mean a scientific or engineering situation in which (a) it is desirable to possess an accurate theory of the processes giving rise to available data, but (b) such a theory cannot be deduced in the strict logical sense from this data. The solution of a discovery problem thus requires some kind of inductive reasoning, and the ability to solve a range of discovery problems requires an inductive method of wide applicability.

Contemporary research within the foregoing framework may be divided into two categories according to the expressiveness of the theories emitted by envisioned systems of inductive inference. The first category deals with theories based on knowledge representations like recursive functions, formal languages, boolean functions, etc. The second deals with more expressive representational systems like first-order languages and extensions thereof. Within each of these categories two research areas have emerged, directed at different models of the data upon which inductive inference is based. In the first of these models, data is made available in some arbitrary order with no assumptions about the statistical processes that govern its generation. In the second model it is assumed that data arise via independent and identically distributed trials with respect to some underlying probability distribution (we refer to this below as *iid* data). We may thus picture the current state of

[1] An entry to the literature is provided in [23, 2].

	Restricted	Expressive
non-*iid* data	I	II
iid data	III	IV

Table 1: Contemporary Research on Machine Inductive Inference

mathematical research on inductive inference as the 2×2 matrix shown in Table 1.

Through the early 1980's most work in machine induction fell in Quadrant I (see [11] for an overview of this research). In 1984, Valiant [27] introduced a new model of inductive inference based on *iid* data. This model relaxed the requirements on the accuracy and reliability of inference algorithms. These relaxed requirements made possible the imposition of more stringent demands concerning efficiency, both in terms of the amount of data examined, and the resources consumed to examine them. Valiant's approach gave rise to the research thrust in Quadrant III which has yielded quantitative results relating the time complexity of learning algorithms to the level of accuracy and reliability demanded of the solutions they provide. Blumer *et al.* [1] elaborated and extended Valiant's model of machine induction to give a deep mathematical analysis of the conditions under which a wide range of discovery problems can be solved within this model. Their analysis has led to a vigorous research effort on the part of many researchers devoted to investigating reliable and efficient inference of classes of geometric concepts, recursive functions and formal languages (see [23]).

Simultaneous with the foregoing developments in Quadrant III, Osherson & Weinstein [14, 16, 18]— building on earlier work of Shapiro [26] and Glymour [6] — introduced a model of inference for first-order logical structures which extended the research in Quadrant I to the realm of highly expressive systems for knowledge representation. This work thus falls into Quadrant II. We obtained general results about the identification of classes of relational structures and about the behavior of algorithms satisfying various computational restrictions. In recent work, we have extended the research thrust in Quadrant III to the quantitative study of algorithms for inferring properties of relational structures. This

latter work thus belongs to Quadrant IV. In order to pursue this study, we have developed mathematical definitions of approximate truth which allow us to extend the *iid* data model to discovery problems which arise in the context of first-order logic. It is our hope that these developments open the way to further quantitative results on machine inductive inference in the domain of highly expressive knowledge representations.

In this paper, we briefly describe our research on inductive inference corresponding to quadrants II and IV. Our work in quadrant II has focussed on a paradigm of scientific discovery known as "truth detection" wherein an inductive agent is responsible for determining the truth value of a first-order sentence in an unknown structure. Within this paradigm, data are presented in arbitrary order. In contrast, our research in quadrant IV has been devoted to articulating concepts of approximate truth and investigating the inference of approximately true theories on the basis of *iid* data drawn from arbitrary relational structures. We now proceed to describe some highlights of this research, beginning with approximate truth.

APPROXIMATE TRUTH

Ideally, we desire our inductive inference agents to provide complete solutions to the problems posed to them, to work with 100% reliability, and to be computationally feasible. It was thus an essential contribution to the theory of learning to discover that for many situations of interest to us, the existence of such inductive inference agents is ruled out in principle. Such was the message of Gold's [7] seminal paper and of much of the research to which it gave rise. Valiant's [27] paper may be viewed as a response to the negative theorems of this literature. He showed that for small sacrifices in reliability and accuracy, efficient inductive inference algorithms could be designed for nontrivial learning problems. Valiant's paradigm came to be known as "Probably Approximately Correct" (or PAC) learning since the solution sought need not be entirely correct nor obtained with perfect reliability.

One goal of our research program has been to exploit Valiant's insight in the context of learning problems involving highly expressive languages, notably, first-order logical languages. For this purpose it is first necessary to formulate a sense in which solutions to such problems can be

partial. Two approaches have been pursued. The first approach extends the PAC framework in a straightforward way to the first order context. The second approach adopts a new analysis of the sense in which a first-order sentence may be approximately true and then investigates algorithms designed to discover approximate truth-values for such sentences in a wide class of potential situations. We give some idea of the results achieved within each approach, starting with our extension of the PAC framework.

Learning First-Order Concepts in the PAC Framework

Within the PAC learning framework (see [1]), a space of points is selected, along with a collection of its subsets called "concepts". One concept, X, is selected arbitrarily, its content being initially unknown to the learner $\mathcal{L}$. Points are then sampled from the space according to a probability distribution that is also unknown to $\mathcal{L}$. Each sampled point is labeled as falling in or out of X. $\mathcal{L}$ must convert the sampled points into a concept X' such that the probability of the symmetric difference of X and X' is low according to the distribution that governs sampling. It is desired that regardless of the concept X that was chosen before sampling began, the probability is high that a sample of points will be drawn that lead $\mathcal{L}$ to a successful conjecture. In this case, the concept-class is said to be "learnable" in the space. We assume familiarity with the quantitative version of this concept-learning paradigm, which is presented in [1]. For simplicity in what follows, we allow learners to be any function from labeled samples to concepts, excluding coin tosses as further inputs.

Now in a practical setting, the set of concepts cannot be arbitrary subsets of the given space. In order to be useful they must at least have finite descriptions in a well-behaved language; otherwise the learner could not communicate her findings to anyone else. First-order logic provides a set of drescriptions of finite character, so we now proceed to embed the foregoing paradigm in a model-theoretic context. Our discussion will be relatively nontechnical.

To begin, we fix an arbitrary, nonlogical vocabulary and denote the resulting predicate calculus (with identity) by $\mathcal{L}$. For example, the nonlogical vocabulary might consist of a single binary relation symbol S. The set of sentences of $\mathcal{L}$ — that is, the formulas in which all variables

are bound – are also denoted by $\mathcal{L}$. Let x denote a distinguished free variable of $\mathcal{L}$. By $\mathcal{L}(x)$ we denote the set of formulas in which just the variable x occurs free. Thus, for the language based solely on S, the following formulas belong to $\mathcal{L}(x)$.

(1) (a) $\exists yz(Szy \wedge Syx)$
(b) $\forall y(x = y \vee Sxy)$
(c) $\forall y(x = y \vee Syx)$

Suppose now that a model $\mathcal{M}$ of $\mathcal{L}$ is given. Such a model consists of a nonempty set $|\mathcal{S}|$ (called $\mathcal{M}$'s *domain*) plus interpretations of the nonlogical vocabulary in that set. For example, $\mathcal{O} = (\omega, <)$ is a model of the language based on S; the domain $|\mathcal{O}|$ of $\mathcal{O}$ is the set $\omega = \{0, 1, 2, \ldots\}$. Each model determines the truth value of every $\theta \in \mathcal{L}$; for example, $\exists x \forall y(x = y \vee Sxy)$ is true in $\mathcal{O}$ and $\exists x \forall y(x = y \vee Syx)$ is false. Similarly, each model assigns a subset of its domain to every $\varphi \in \mathcal{L}(x)$; this set consists of exactly the domain elements a such that φ is true in the model when x is interpreted as a. To illustrate, $\mathcal{O}$ assigns the sets $\{2, 3, \ldots\}$, $\{0\}$, and $\emptyset$ to (1)a,b,c, respectively. It may thus be seen that any pair $(\mathcal{M}, \Phi)$ consisting of a model $\mathcal{M}$ for $\mathcal{L}$ and a subset Φ of $\mathcal{L}(x)$ determines a concept-learning problem of the PAC variety. For example, $\mathcal{O}$ and (1) determine the problem in which ω is the underlying space of points, and the extensions of (1)a,b,c in $\mathcal{O}$ are the collection of concepts.

Given a class $\mathcal{K}$ of models and $\Phi \subseteq \mathcal{L}(x)$, Φ is said to be *learnable in* $\mathcal{K}$ just in case Φ is PAC-learnable in every $\mathcal{S} \in \mathcal{K}$. Within this analysis two mathematical problems arise. They may be stated as follows.

(2) (a) Given a set $\Phi \subseteq \mathcal{L}(x)$, characterize the models in which Φ is learnable, and the models in which Φ is not learnable.
(b) Given a collection $\mathcal{K}$ of models, characterize the sets of formulas that can be learned in $\mathcal{K}$, and the sets of formulas that cannot be learned in $\mathcal{K}$.

To address these questions, a fundamental tool is the work of Blumer et al. [1] relating VC-dimension to learnability. Relying on their results, we have been able to prove a variety of theorems bearing on (2)a,b. One finding of a positive character followed by one of a negative character may be described here; details, proofs, and further results are provided

in [19]. The following standard terminology will be helpful. A set $T \subseteq \mathcal{L}$ is called a *theory*. Given theory T and model $\mathcal{S}$, we write $\mathcal{S} \models T$ just in case every member of T is true in $\mathcal{S}$.

First finding: A theory T is called *strong* just in case it meets the following conditions, for all models $\mathcal{S}, \mathcal{U}$:

(a) if $\mathcal{S} \models T$ then $|\mathcal{S}|$ is infinite;

(b) if $\mathcal{S} \models T, \mathcal{U} \models T$, and both $\mathcal{S}$ and $\mathcal{U}$ have denumerable domains, then $\mathcal{S}$ and $\mathcal{U}$ are isomorphic (in other words, T is "ω-categorical").

For example, the theory of dense linear orders without end points is strong (see [3, Proposition 1.4.2]). The following theorem shows that the class of all first-order concepts can be learned in any model of a strong theory.

(3) THEOREM: Suppose that T is a strong theory. Then $\mathcal{L}(x)$ is learnable in $\{\mathcal{S} \mid \mathcal{S} \models T\}$.

Second finding: Given a set $\Phi \subseteq \mathcal{L}(x)$, we say that a theory T *expresses the learnability of* Φ just in case for all models $\mathcal{S}$, Φ is learnable in $\mathcal{S}$ iff $\mathcal{S} \models T$. Such theories have the useful property of providing a test for learnability in given situations. Unfortunately, no theory expresses the learnability of even relatively simple subsets of $\mathcal{L}(x)$. This is the content of the next theorem, stated with the following notation. The subset of $\mathcal{L}(x)$ of form $\exists y \forall z \varphi(xyz)$, with φ quantifier-free is denoted by $\mathcal{L}_{\exists\forall}(x)$.

(4) THEOREM: Suppose that $\mathcal{L}$ contains at least one binary relation symbol. Then there is no theory that expresses the learnability of $\mathcal{L}_{\exists\forall}(x)$.

Determining the Approximate Truth of First-Order Theories

Our second approach to Discovery Problems within Quadrant IV starts from a definition of the concept "first order sentence θ is approximately true in relational structure $\mathcal{S}$." We shall here limit ourselves to brief discussion of this idea; details are provided in [12, 13]. Our theory starts from consideration of the degree to which one structure approximates another. Approximate truth in a structure is then construed as (exact) truth in an approximating structure. It is not claimed that this approach illuminates every aspect of the problem of approximate truth. Rather, our theory is designed for situations of the following kind.

Let us conceive of a narrow strip of land (e.g., a coastline) undergoing mineral exploration. A point along the strip is to be designated randomly according to some unknown probability distribution. Once the site is designated, it will be decided whether to drill at that location. Let p be a variable for points along the strip, and consider the following predicates and hypothesis (5).

$Lp \equiv$ a lode exists within 1000 feet of the surface at point p.

$Rp \equiv$ there is superficial igneous rock at point p.

(5) $(\forall p)(Lp \rightarrow Rp)$

Even if false about the actual strip under exploration, (5) might be useful if true about a fictitious strip that approximates it. In this case, (5) can be considered to be approximately true about the actual strip.

To give substance to the foregoing idea, let the actual and fictitious strips be represented by the same, real interval I. Let $\mathbf{L}$, $\mathbf{R}$ be the extensions of L and R in the actual strip, and $\mathbf{L}'$, $\mathbf{R}'$ be their extensions in the fictitious strip. For the fictitious strip to approximate the actual one we require that every point in $\mathbf{L}'$ be near to some point in $\mathbf{L}$, and that every point in the complement of $\mathbf{L}'$ be near to some point in the complement of $\mathbf{L}$; similarly for $\mathbf{R}'$ and $\mathbf{R}$. It is natural, however, to ask for greater nearness in high probability subregions than in low probability subregions since our judgment about drilling is more likely to be put to the test in the former than in the latter. We thus define the "probability distance" of two points to be the probability mass of the interval that separates them. It can be seen that two points separated by a small

probability distance are either metrically close in a high mass interval or else common members of a low mass interval.

Now fix $b \in (0,1)$. The fictitious strip is called a "b-variant" of the actual strip just in case for every point p' there is a point p such that p' is within probability distance b of p, and $p' \in \mathbf{L}'$ iff $p \in \mathbf{L}$; likewise for $\mathbf{R}'$ and $\mathbf{R}$. Thus, for the fictitious strip to be a b-variant of the actual one, every point $p' \in \mathbf{L}'$ must be justified by a nearby point $p \in \mathbf{L}$; likewise, every point $p' \notin \mathbf{L}'$ must be justified by a nearby point $p \notin \mathbf{L}$ —and similarly for $\mathbf{R}'$ and $\mathbf{R}$. In this case, we consider the fictitious strip to approximate the actual one, up to the parameter b. Sentences like (5) are considered to be "b-true" in the actual strip just in case they are standardly true in at least one of its b-variants.

The following example illustrates the potential usefulness of b-true sentences. Let I, $\mathbf{L}$, and $\mathbf{R}$ be as described above. We imagine that I has been partitioned into ten regions. A point will be drawn randomly from I according to unknown, continuous probability distribution P, and the following question will be posed.

(6) $p \in \mathbf{L}$ for all p in the region from which the sampled point was drawn?

Suppose that inspection reveals there to be no superficial igneous rock in the region actually sampled, and that hypothesis (5) is known to be .01-true in the strip. Then, it may be proved that (6) is false with probability at least .80.

Our theory is a generalization of the foregoing illustration. We have pursued its development from both the deductive and inductive points of view. For brevity, only the inductive logic of approximate truth will be considered here. Our approach is based on a paradigm of empirical inquiry that may be called "probably approximately correct truth detection." Within this paradigm scientists convert a given first-order sentence θ along with accuracy and reliability parameters b,c into a set of queries. The queries bear on the interpretation of predicates within a fixed but unknown structure $\mathcal{S}$. Illustrating with a unary predicate A, these queries take the form:

> Does the nth randomly sampled point from the domain of $\mathcal{S}$ fall into the $\mathcal{S}$-extension of A?

The scientist has no knowledge of the probability distribution that governs random sampling from $\mathcal{S}$'s domain. On the basis of a set of queries whose size grows no faster than polynomially in $\frac{1}{b}$ and $\frac{1}{c}$ the scientist must emit, with reliability at least $1-c$, a b-truth-value for θ in $\mathcal{S}$. In [12] we show that there is a formal scientist that succeeds in this task with respect to a wide class of first-order sentences and structures. Specifically, the class of sentences for which our method is provably successful includes all sentences in which no predicate letter occurs both negatively and positively. Such sentences are called "monotone." The class of structures in which the scientist can infer an approximate truth-value for monotone sentences includes all structures with continuous domain and measurable extensions for predicates. Placed in an arbitrary and unknown member of this extensive class of structures, and parameterized by any monotone sentence θ and any choice of accuracy and reliability parameters b,c, the scientist we define makes only polynomially many queries in $\frac{1}{b}$ and $\frac{1}{c}$ and emits, with reliability at least $1-c$, a b-truth-value for θ in the unknown structure.

Our current work on this topic is devoted to extending the foregoing result to nonmonotone sentences, to exploring alternative conceptions of approximate truth, and to investigating other paradigms of scientific discovery in which approximate truth is a satisfactory goal.

TRUTH DETECTION

By a paradigm of scientific discovery let us understand any specification of the concepts "scientist" and "discovery problem" along with a criterion that determines the conditions under which a given scientist is credited with solving a given problem. In our research in quadrant II we have investigated several paradigms of scientific discovery in which discovery problems are characterized using first-order logical languages and extensions thereof. We here describe one such paradigm, truth detection, and some of the recent results obtained about it.

Let a countable, first-order language $\mathcal{L}$ with identity be fixed, suitable for expressing scientific theories and data in some field of empirical inquiry. Prior research in the field is conceived as verifying a set T of $\mathcal{L}$-sentences, which constitute the axioms of a theory already known to be true. Each model of T thus represents a possible world consistent

with background knowledge. Nature has chosen one of these models – say, structure $\mathcal{S}$ – to be actual; her choice is unknown to us. (For ease of exposition, we will suppose that Nature's choice is limited to countable models of T.)

Scientisits are conceived as attempting to divine the truth-value in $\mathcal{S}$ of specific sentences not decided by T. Suppose that scientist Ψ wishes to determine the truth-value of θ in $\mathcal{S}$. At the start of inquiry, Ψ knows no more about $\mathcal{S}$ than what is implied by T. As inquiry proceeds, more and more information about $\mathcal{S}$ becomes available. This information has the following character. We conceive of Ψ as being able to determine, for each atomic formula φ of $\mathcal{L}$ and any given sequence of objects from the universe of $\mathcal{S}$ whether or not that sequence satisfies φ in $\mathcal{S}$. Ψ receives the entire universe of $\mathcal{S}$ in piecemeal fashion and bases its conjecture at a given moment on the finite subset of the universe of $\mathcal{S}$ examined by that time. In response to each new datum, Ψ emits a fresh conjecture about the truth of θ in $\mathcal{S}$, announcing either "true" or "false." To be counted as successful, Ψ's conjectures must stabilize to the correct one. Notice that no assumption is made about the process generating the data Ψ receives; in particular, in order to successfully detect the truth of θ in $\mathcal{S}$, we require Ψ to stabilize to a correct conjecture no matter what data sequence is presented. (This distinguishes the current model from the *iid* data model presented in the preceding section where data sequences are generated by randomly sampling points from a structure according to some time invariant probability distribution over the universe of that structure.)

Let us summarize the above discussion with the following definition: we say a scientist Ψ *detects the truth-value of* a sentence θ *with respect to* background knowledge T just in case for every countable model $\mathcal{S}$ of T and every data sequence e generated from $\mathcal{S}$, Ψ stabilizes to a correct conjecture about the truth-value of θ in $\mathcal{S}$.

Our research on truth detection has addressed the following questions, among others:

- For which sentences θ and theories T do there exist scientists who detect the truth of θ with respect to T?
- Are there theories T such that some single scientist detects the truth-value of all sentences with respect to T?

- How are the answers to the preceding questions altered if we impose computational or methodological constraints on the scientists in question?

We have also examined a further question, the answer to which provides considerable information about the choice of first-order logic as a mode of representation for background knowledge. The remainder of this section is devoted to this matter.

With respect to the paradigm of truth detection, we may view discovery problems as parameterized by a theory T which represents the background knowledge available to a scientist at the outset of investigation. When discovery problems are so viewed the following uniformity question naturally presents itself. Is there a uniform method M for solving the problem posed by T, if that problem is solvable at all? Such a method M might be uniform in T in the following sense. In the course of computing its conjecture about the truth-value of the sentence θ, M could receive answers to any queries it chose about the membership of individual sentences in T. M's computation of its conjecture in the face of incoming data would then be entirely effective relative to the answers it received to its queries. Such a method M may be represented by a Turing machine with oracle. If M is such an oracle machine, we write M^T to denote the scientist computed by M when equipped with an oracle for T. Given this understanding of uniform method for the solution of discovery problems, the following theorem provides an affirmative answer to the above question.

(7) THEOREM: There is an oracle machine M such that for all sets of first-order sentences T and all first-order sentences θ, if there is a scientist who detects the truth-value of θ with respect to T, then M^T detects the truth-value of θ with respect to T.

Proof of the theorem is provided in [17].

Theorem (7) leads to a fundamental question about the role of first-order logic in inductive inference. We ask: In making possible the uniform solution of scientific discovery problems, what is the role of our choice of first-order logic for representing background knowledge? Could some yet more expressive logical language be used to represent background knowledge that would still allow for a uniform solution of the

discovery problems thus represented? Our research has provided a partial answer to these questions. In order to give the answer, we will need to consider a slight strengthening of the paradigm of truth detection and introduce some concepts from the theory of models.

Let $\mathcal{L}'$ be a regular logical language which contains our first-order language $\mathcal{L}$. (For the concept *regular logical language* see [4]; suffice it to say that first-order logic itself is a regular logical language as are most examples of extensions of first-order logic, such as second-order logic, extensions by the addition of cardinality quantifiers, etc.) Let T be a set of $\mathcal{L}'$ sentences, θ a sentence of $\mathcal{L}$, and $\mathcal{S}$ a model of T. We say that e is a restricted data sequence for $\mathcal{S}$ and θ just in case e is the result of removing all information about atomic formulas containing vocabulary not present in θ from some data sequence e' generated from $\mathcal{S}$. We say that scientist Ψ *strongly detects the truth-value of* θ *with respect to* T just in case for every countable model $\mathcal{S}$ of T and every restricted data sequence e for $\mathcal{S}$ and θ, Ψ stabilizes to a correct conjecture about the truth-value of θ in $\mathcal{S}$. Finally, we say that $\mathcal{L}'$ has the *uniform strong detection property* just in case there is an oracle machine M such that for all sets of $\mathcal{L}'$-sentences T and all first-order sentences θ, if there is a scientist who strongly detects the truth-value of θ with respect to T, then M^T strongly detects the truth-value of θ with respect to T. Our preceding theorem may be strengthened to exhibit further uniformity in the solvability of discovery problems characterized by first-order theories as follows.

(8) THEOREM: First order logic has the uniform strong detection property.

We are now in a position to show the extent to which first-order logic is unique in affording uniform solution of discovery problems. A regular logical language $\mathcal{L}'$ has the *Lowenheim-Skolem property* just in case every satisfiable $\mathcal{L}'$-sentence has a countable model. It is a fundamental fact of model theory that first-order logic has the Lowenheim-Skolem property. The following theorem provides a characterization of first-order logic as a maximal regular logic with the Lowenheim-Skolem and uniform strong detection properties (see [17] for proof).

(9) THEOREM: Let $\mathcal{L}'$ be a regular logical language containing first-order logic $\mathcal{L}$. If $\mathcal{L}'$ has the Lowenheim-Skolem property and the

uniform strong detection property then $\mathcal{L}' = \mathcal{L}$.

Theorem (9) indicates that first-order logic has a special status as a knowledge representation language for scientific discovery problems. This result also suggests important topics for further research. First, are there proper extensions of first-order logic which fail to have the Lowenheim-Skolem property but do allow for uniform solution of discovery problems? Second, are there languages whose expressive power is incomparable with that of first-order logic which allow for uniform solution of discovery problems? Such languages might arise as fragments of proper extensions of first-order logic. The answers to both these questions may have significance for the choice of knowledge representation languages for discovery problems which arise in scientific or technological contexts. We plan to investigate these and related issues in our continuing research on automated scientific discovery.

CONCLUDING REMARKS

Each paradigm of empirical inquiry studied within Computational Learning Theory is a mathematical abstraction from the complex web of issues indicated in the introduction above. Study of these models is aimed at faciliating the development of practical algorithms for the automated solution of discovery problems arising in practice. It may also be hoped, as well, that results within the theory partially clarify some of the questions that surround the nature of scientific activity itself. Some of our work has been focussed on such questions (e.g., [5, 15, 21]). The present discussion concludes with a brief summary of one pertinent result.

Scientific inference is an essentially non-deductive affair inasmuch as true theories — apart from trivial, exceptional cases – cannot be deduced from the data available to scientists. Nonetheless, deductive logic is widely recognized to play a central role in scientific thought, for example, in drawing out the consequences of a theory for empirical test. For this reason deductive logic has been central to the analysis of several components of scientific activity. To illustrate, it has been suggested that the confirmation of a scientific theory is a function of the empirical verification of its logical consequences (see [8] for discussion).

Unfortunately, a simple analysis of confirmation on this basis founders on the richness of the set of logical consequences of a given theory. Thus, one consequence of the axiom A is $A \vee S$ for arbitrary sentence S; yet verification of S (hence of $A \vee S$) need not confirm A.

To save the insight behind the idea that confirmation of consequences yields inductive support, it is tempting to exclude inferences like $A \models A \vee S$ from the set of "scientifically relevant" deductions. After all, this latter inference has a suspicious character inasmuch as it does not depend on any particular relation between A and S. Following this line of thought, several definitions of scientifically relevant deduction have been advanced, leading to fruitful analyses of confirmation and theory-comparison (see [29, 28, 24, 25]). To be pertinent to our understanding of scientific practice, however, a definition of relevant deduction must satisfy a further criterion. It must be the case that scientists whose deductive reasoning is limited to relevant inferences are just as scientifically competent as scientists not so limited. That is, for every scientific problem that is solvable in principle, there must exist a scientist who never reasons in deductively irrelevant fashion yet who also succeeds in solving that problem. Otherwise, the proposed definition of relevant deduction does not allow us to fully understand how it is that science sometimes succeeds.

Starting with a simple definition of relevant deduction due to Schurz & Weingartner [25] we have shown that for every solvable problem of the kind described in the last section there is indeed a successful scientist whose deductive reasoning conforms to the definition. Details are given in [22]. Evidence is thereby provided that the kind of definition proposed in Schurz & Weingartner [25] is plausible as a representation of the deductive component of scientific reasoning. In this way, study of formally defined paradigms of inductive inference within Computational Learning Theory can shed some light on the foundations of scientific inquiry.

References

[1] Blumer, A., A. Ehrenfeucht, D. Haussler, & M. Warmuth (1987). *Learnability and the Vapnik-Chervonenkis Dimension* (Technical Report UCSC-CRL-87-20). Santa Cruz: University of California.

[2] Case, J. & Fulk, M. (Eds.) (1990). *Proceedings of the third annual workshop on computational learning theory*, San Mateo, CA: Morgan-Kaufmann.

[3] Chang, C. C. & Keisler, H. J. (1973). *Model theory*, Amsterdam: North-Holland.

[4] Ebbinghaus, H.-D. (1985). Extended logics: the general framework. In Barwise, J. & S. Feferman (eds.), *Model theoretic logics*. New York: Springer-Verlag.

[5] Gaifman, H., Osherson, D. & Weinstein, S. (1990). A reason for theoretical terms. *Erkenntnis, 32,* 149-159.

[6] Glymour, C. (1985). Inductive inference in the limit. *Erkenntnis, 22,* 23-31.

[7] Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10,* 447-474.

[8] Hempel, C. G. (1965) *Aspects of scientific explanation and other essays in the philosophy of science*. The Free Press.

[9] Langley, P., Bradshaw, G., & Simon, H. (1983). Rediscovering chemistry with the BACON system. In R. Michalski, J. Carbonell, & T. Mitchell (Eds.) *Machine learning: An artificial intelligence approach*. Palo Alto, CA: Tioga.

[10] Langley, P. & Nordhausen, B. (1986). A framework for empirical discovery. In *Proceedings of the International Meeting on Advances in Learning,* Les Arcs, France.

[11] Osherson, D., Stob, M., & Weinstein, S. (1986). *Systems that Learn*. Cambridge, MA: MIT Press.

[12] Osherson, D., Stob, M., & Weinstein, S. (1989). On approximate truth. In R. Rivest, D. Haussler, & M. Warmuth (Eds.), *Proceedings of the second annual workshop on computational learning theory*. San Mateo, CA: Morgan-Kaufmann.

[13] Osherson, D., Stob, M., & Weinstein, S. (1989). *A theory of approximate truth*. (Technical Report). Cambridge, MA: M.I.T.

[14] Osherson, D. & Weinstein, S. (1986). Identification in the limit of first-order structures. *Journal of Philosophical Logic, 15,* 55-81.

[15] Osherson, D., Stob, M., & Weinstein, S. (1988). Mechanical learners pay a price for Bayesianism. *Journal of Symbolic Logic, 53,* 1245-1251.

[16] Osherson, D. & Weinstein, S. (1989). Paradigms of truth detection. *Journal of Philosophical Logic, 18,* 1-42.

[17] Osherson, D., Stob, M., & Weinstein, S. (1991). A universal inductive inference machine," *Journal of Symbolic Logic, 56,* 661-672.

[18] Osherson, D., Stob, M., & Weinstein, S., (in press). A universal method of scientific inquiry. *Machine Learning.*

[19] Osherson, D., Stob, M., & Weinstein, S. (1991). New directions in automated scientific discovery. *Information Sciences.*

[20] Osherson, D. & Weinstein, S. (1989). Identifiable collections of countable structures. *Philosophy of Science, 56,* 95-105.

[21] Osherson, D. & Weinstein, S. (1990). On advancing simple hypotheses. *Philosophy of Science, 57,* 266-277.

[22] Osherson, D. & Weinstein, S. (in press). Relevant consequence and scientific discovery. *Journal of Philosophical Logic.*

[23] Rivest, R., Haussler, D., & Warmuth, M. (Eds.) (1989). *Proceedings of the second annual workshop on computational learning theory.* San Mateo, CA: Morgan-Kaufmann.

[24] Schurz, G. (1991). Relevant deduction. *Erkenntnis.*

[25] Schurz, G. & Weingartner, P. (1987). Verisimilitude defined by relevant consequence-elements: A new reconstruction of Popper's idea. In T. A. Kuipers (Ed.), *What is closer-to-the-truth?* Amsterdam: Rodopi.

[26] Shapiro, E. (1981). An algorithm that infers theories from facts. In *Proceedings of the seventh international joint conference on artificial intelligence.*

[27] Valiant, L. (1984). A theory of the learnable. *Communications of the ACM, 27,* 1134-1142.

[28] Weingartner, P. (1988). Remarks on the consequence-class of theories. In E. Scheibe (Ed.), *The role of experience in science.* Walter de Gruyter.

[29] Weingartner, P. & Schurz, G. (1986). Paradoxes solved by simple relevance criteria. *Logique et Analyse.*

[30] J. Zytkow (1987). Combining many searches in the FAHRENHEIT discovery system. In *Proceedings of the Fourth International Workshop on Machine Learning,* Irvine CA.