Logic and Learning

MS-CIS-90-65
LOGIC & COMPUTATION 24

Daniel N. Osherson
M. I .T.

Michael Stob
Calvin College

Scott Weinstein
University of Pennsylvania

Department of Computer and Information Science
School of Engineering and Applied Science
University of Pennsylvania
Philadelphia, PA 19104

September 1990

Logic and Learning*

Daniel N. Osherson
M. I. T.

Michael Stob
Calvin College

Scott Weinstein
University of Pennsylvania

September 11, 1990

Abstract

The theory of first-order logic — or "Model Theory" — appears in few studies of learning and scientific discovery. We speculate about the reasons for this omission, and then argue for the utility of Model Theory in the analysis and design of automated systems of scientific discovery. One scientific task is treated from this perspective in detail, namely, concept discovery. Two formal paradigms bearing on this problem are presented and investigated using the tools of logical theory. One paradigm bears on PAC learning, the other on identification in the limit.

1 Introduction

The predicate calculus provides a convenient medium for expressing facts and hypotheses, and it is thus no suprise that numerous systems of machine learning are designed to discover predicate logic sentences that summarize or extend the data presented to them (e.g., [16]). Theoreticians of learning have also found the language of logic to be of central concern, as witnessed by influential studies that bear on the discovery of various kinds of formulas (e.g., [11, 13, 24, 14, 1]). In contrast, the theory of the predicate calculus — that is, contemporary Model Theory (see, e.g., [5]) — rarely emerges in theoretical studies of learning, at least within the movement represented by [12, 22, 9]. We may speculate about two reasons for this absence. First, other mathematical theories — notably, the theories of computation, complexity and probability — have yielded a rich harvest of results, so it is natural that researchers continue to focus on these tools in their analysis of learning. Second, deductive logic seems to be divorced from the inductive processes that lie at the heart of learning, since the inferences involved in empirical discovery are uncertain and subject to retraction, which is quite the opposite of deductive inference.

*Research support was provided by the Office of Naval Research under contract No. N00014-87-K-0401 to Osherson and Weinstein, and by a Siemens Corporation grant to Osherson. Correspondence to D. Osherson, E10-044, M.I.T., Cambridge, MA 02139; e-mail: dan@psyche.mit.edu

The purpose of the present paper is to suggest that, appearances notwithstanding, Model Theory is a potentially valuable tool for understanding learning algorithms designed to discover predicate logic sentences. As evidence for this suggestion, we consider the problem of producing necessary and sufficient conditions for a concept whose extension is available in part or in whole. Two model-theoretic perspectives are proposed; the first is related to PAC learning in the sense of [26, 3], the second to identification in the limit in the sense of [10, 2]. As a preliminary, we observe that Model Theory need not be viewed as bearing primarily on deductive implication. Rather, implication may be seen as derivative to the primary concern of the theory, namely, the conditions under which specified sentences are true in given situations (or "models"). It is the concern for truth-in-a-situation that renders the theorems of Model Theory relevant to discovering a true description of one's environment.[1]

2 Learning first-order concepts in the PAC framework

The present section considers the following situation, studied within PAC learning (see [3, 7]). A space of points is selected, along with a collection of its subsets (called "concepts"). One of the concepts, C, is arbitrarily selected, and points are sampled from the space according to an unknown probability distribution. Each sampled point is labeled as falling in or out of C. The learner must convert this information into a conjectured concept C' such that the probability of the symmetric difference of C and C' is low according to the unknown distribution that governs sampling. It is desired that regardless of the concept chosen, there is a high probability of drawing a small sample of points leading the learner 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, as presented, for example, in [3]. For simplicity in what follows, we allow learners to be any function from labeled samples to concepts, excluding coin tosses as further inputs.

In a practical setting, the set of concepts cannot be arbitrary subsets of the given space. At the least, they must have finite descriptions in a well-behaved language since otherwise the learner could not communicate his findings to anyone else. First-order logic provides descriptions of the required character, and we now proceed to embed the foregoing paradigm in a model-theoretic context. Our discussion will be brief and 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 R. The sentences of $\mathcal{L}$ — i.e., the formulas without free variables – 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 R, the following formulas belong to $\mathcal{L}(x)$.

(1) (a) $\forall y(x = y \vee Rxy)$

(b) $\forall y(x = y \vee Ryx)$

[1] Formulas of the predicate calculus are often called "first-order" to distinguish them from formulas with more complex kinds of quantification. We sometimes employ this terminology in what follows.

(c) $\exists yz(Rzy \wedge Ryx)$

Suppose now that a model $\mathcal{S}$ of $\mathcal{L}$ is given. Such a model consists of a nonempty set $|\mathcal{S}|$ (called $\mathcal{S}$'s *domain*) along with interpretations of the nonlogical vocabulary in that set. For example, $\mathcal{O} = (\omega, <)$ is a model of the language based on R; 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, $\exists x \forall y(x = y \vee Rxy)$ is true in $\mathcal{O}$ and $\exists x \forall y(x = y \vee Ryx)$ is false. Similarly, each model assigns a subset of its domain to every $\varphi \in \mathcal{L}(x)$, namely, the set of domain elements a such that φ is true in the model when x is interpreted as a. To illustrate, $\mathcal{O}$ assigns the sets $\{0\}$, $\emptyset$, and $\{2, 3, \ldots\}$ to (1)a,b,c, respectively. It may thus be seen that any pair $(\mathcal{S}, \Phi)$ consisting of a model $\mathcal{S}$ 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 $\mathcal{S} \in \mathcal{K}$. The principal problems that arise in this context are 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 $\mathcal{K}$.

A fundamental tool for addressing these problems is the work of Blumer, Haussler, Ehrenfeucht and Warmuth [3] relating VC-dimension to learnability. Relying on their results, we have been able to prove several 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 [5, Proposition 1.4.2]). The following theorem says 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 are useful inasmuch as they provide 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 $\mathcal{L}_{\exists\forall}(x)$.

3 Discovering first-order intensions in the limit

3.1 Overview

The present section is devoted to a paradigm in which the entire extension of a target concept is revealed to the scientist in piecemeal fashion. In response to these data, the scientist advances a succession of first-order formulas in the hope of stabilizing on a necessary and sufficient condition for membership in the concept. For notational convenience we consider only unary concepts; extension to concepts of arbitrary arity is straightforward. The paradigm is formalized in the present section. Section 4 is devoted to theorems. A related paradigm is studied in [20], and several proofs below will refer to constructions appearing there. On the other hand, new techniques are used to prove the results of Sections 4.2 and 4.3, and they illustrate the use of model-theoretical constructions in the study of learning.

3.2 Paradigm

3.2.1 Language and models

We fix a countable, first-order language $\mathcal{L}$ (with identity) that includes a distinguished, unary predicate C.[2] This predicate represents the target concept for which a first-order intension is sought. We also distinguish a variable x and denote by $\mathcal{L}(x)$ the set of all formulas of $\mathcal{L}$ in which just the variable x occurs free, and in which C does not occur. $\mathcal{L}(x)$ represents the set of potential intensions for the target concept; C is excluded from the vocabulary of $\mathcal{L}(x)$ in order to rule out intensions that are accurate but trivial (e.g., the formula Cx).

A formula φ of $\mathcal{L}$ is *basic* just in case φ is an atomic formula or the negation of such. The set of all basic formulas is denoted BAS.

We conceive of Nature as choosing one member from a class $\mathcal{K}$ of models of $\mathcal{L}$. $\mathcal{K}$ is conceived as representing the class of "possible worlds" known to the scientist to be theoretical alternatives prior to his inquiry. Attention is limited to models with countable domains.

[2]The countability of $\mathcal{L}$ means that $\mathcal{L}$'s vocabulary is countable and that $\mathcal{L}$ includes denumerably many individual variables.

Henceforth, by "model" we understand "countable model that interprets $\mathcal{L}$." By a *complete assignment* to a model $\mathcal{S}$ is meant any mapping of the (countable) set of variables of $\mathcal{L}$ onto $|\mathcal{S}|$. Thus, a complete assignment to $\mathcal{S}$ provides every member of its domain with at least one temporary name.

3.2.2 The data made available to scientists

An *environment* is any ω-sequence over BAS.[3] The set of formulas appearing in an environment e is denoted by *range*(e). The initial finite sequence of length $i \in \omega$ in e is denoted $\bar{e}_i$. $A\bar{x} \in$ *range*(e) [respectively, $\neg A\bar{x} \in$ *range*(e)] may be understood as a message from Nature of the form: "The objects assigned temporary names $\bar{x}$ fall [do not fall] into the set that interprets A." The following definition specifies the sense in which a model underlies an environment.

(5) DEFINITION: Let environment e, model $\mathcal{S}$, and complete assignment g to $\mathcal{S}$ be given. e is *for* $\mathcal{S}$ *via* g just in case *range*$(e) = \{\beta \in \text{BAS} \mid \mathcal{S} \models \beta[g]\}$. e is *for* $\mathcal{S}$ just in case e is for $\mathcal{S}$ via some complete assignment.

To illustrate, suppose that the following environment e is for model $\mathcal{S}$.

$$Tx_3 \quad \neg Qx_3x_2 \quad x_4 = x_5 \quad Tx_4 \quad \ldots$$

Then e may be construed as the following, endless message about $\mathcal{S}$ (where we write $P^{\mathcal{S}}$ to denote the set that interprets the predicate P in $\mathcal{S}$).

> The object given temporary name x_3 belongs to $T^{\mathcal{S}}$. The object with temporary name x_3 is such that the pair x_3, x_2 belongs to the complement of $Q^{\mathcal{S}}$. The objects given temporary names x_4 and x_5 are identical. Object x_4 (and hence object x_5) belongs to $T^{\mathcal{S}}$ ..."

Models are determined by their environments. This is the content of the following lemma, proved in [17].

(6) LEMMA: Let environment e and models $\mathcal{S}$ and $\mathcal{U}$ be given. If e is for both $\mathcal{S}$ and $\mathcal{U}$ then $\mathcal{S}$ and $\mathcal{U}$ are isomorphic.

3.2.3 Scientists and success

Scientists are conceived as working in an environment e for a model $\mathcal{S}$ by examining the $\bar{e}_i$ in turn. The scientist announces at each stage some $\varphi \in \mathcal{L}(x)$ to express the hypothesis that $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$.[4] Lemma (6) ensures that no ambiguity arises about the truth of such hypotheses. To proceed formally, let SEQ be the set of all finite sequences over BAS. (Thus,

[3] An ω-sequence over a set X may be conceived as an infinite list $x_1, x_2, \ldots$ of elements drawn from X.

[4] Recall that if $\varphi \in \mathcal{L}(x)$, then x is the only variable occurring free in φ.

SEQ $= \{\bar{e}_i \mid i \in \omega$ and e is an environment$\}$). By a *(formal) scientist* is meant any function from BAS to $\mathcal{L}(x)$. Note that scientists can be computable or uncomputable, total or partial.

To be successful in a given environment, we stipulate that a scientist's successive conjectures must eventually stabilize to a formula that gives an accurate necessary and sufficient condition for membership in the concept expressed by C.

(7) DEFINITION: Let collection $\mathcal{K}$ of models, model $\mathcal{S}$, environment e for $\mathcal{S}$, and scientist Ψ be given.

(a) Ψ *solves* e just in case there is $\varphi \in \mathcal{L}(x)$ such that $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$, and $\Psi(\bar{e}_i) = \varphi$ for all but finitely many $i \in \omega$.

(b) Ψ *solves* $\mathcal{S}$ just in case Ψ solves every environment for $\mathcal{S}$.

(c) Ψ *solves* $\mathcal{K}$ just in case Ψ solves every $\mathcal{S} \in \mathcal{K}$. In this case, $\mathcal{K}$ is *solvable.*

3.3 Examples

We give an example of solvability followed by an example of unsolvability.

3.3.1 Solvability

(8) EXAMPLE: Suppose that $\mathcal{L}$ is limited to the binary relation symbol R plus the distinguished predicate C. Let P be the set of positive integers, N the set of negative integers. The symbol $<$ denotes the usual ordering on all integers. Let $\mathcal{K}$ consist of all models of the either of the forms:

(a) $(P, <, X)$, where $<$ interprets R, and X is a finite or cofinite subset of P that interprets C;

(b) $(N, <, X)$, where $<$ interprets R, and X is a finite or cofinite subset of N that interprets C;

Then $\mathcal{K}$ is solvable.

Proof: We give an informal description of a scientist Ψ that solves $\mathcal{K}$. Ψ is equipped with an enumeration of triples $(y, \mathcal{S}, \varphi)$ such that y is a variable, $\mathcal{S} \in \mathcal{K}$, and $\varphi \in \mathcal{L}(x)$. At each stage in the examination of the environment e, Ψ finds the first triple, $(y, \mathcal{S}, \varphi)$, in the enumeration consistent with the hypotheses:

(a) e is for $\mathcal{S}$,

(b) y is the temporary name of 0 in $\mathcal{S}$, and

(c) $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$.

Ψ then conjectures φ. It is clear that e will cause Ψ to abandon any triple $(y, \mathcal{S}, \varphi)$ such that e is not for $\mathcal{S}$. On the basis of this observation, it is easy to verify that Ψ solves $\mathcal{K}$. ■

3.3.2 Unsolvability

Let $\mathcal{S} = (S, r_1, r_2, \ldots, X)$ be a model for $\mathcal{L}$, where X interprets C. If $|\mathcal{S}|$ is infinite then there are uncountably many choices for X. On the other hand, there are only countably many formulas in $\mathcal{L}(x)$. Consequently, for fixed $r_1, r_2, \ldots$, the collection $\mathcal{K} = \{(S, r_1, r_2, \ldots, X) \mid X \subseteq S\}$ is trivially unsolvable inasmuchas necessary and sufficient conditions for membership in C cannot be expressed for some choices of X. A nontrivial example of unsolvability is given next. Its verification depends on the following lemma. The set of variables appearing in a given $\sigma \in$ SEQ is denoted by $var(\sigma)$; the conjunction of the members of σ is denoted by $\bigwedge\sigma$.

(9) LEMMA: Let scientist Ψ, and model $\mathcal{S}$ be given. Suppose that Ψ solves $\mathcal{S}$. Then there is $\sigma \in$ SEQ, $p : var(\sigma) \to |\mathcal{S}|$, and $\varphi \in \mathcal{L}(x)$ such that:

(a) $\mathcal{S} \models \bigwedge\sigma[p]$;

(b) $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$.

(c) for all $\gamma \in$ SEQ, if

i. $\sigma \subseteq \gamma$ and

ii. $\mathcal{S} \models \exists x_1 \ldots x_k \bigwedge\gamma[p]$, where $var(\gamma) - var(\sigma) = \{x_1 \ldots x_k\}$

then $\Psi(\gamma) = \varphi$.

The proof of Lemma (9) is easily adapted from a similar result proved in [20, Lemma 27].[5]

(10) EXAMPLE: Suppose that $\mathcal{L}$ is limited to the binary relation symbol R plus the distinguished predicate C. Let $\omega + \omega$ represent two copies of the natural numbers ordered this way: $0, 1, 2, \ldots, 0, 1, 2, \ldots$. The symbol $<$ denotes the usual ordering on ω or $\omega + \omega$. Let $\mathcal{K}$ consists of all models of the either of the forms:

(a) $(\omega, <, \{i\})$, where $<$ interprets R, and $i \in \omega$.

(b) $(\omega + \omega, <, \{\underline{0}\})$, where $<$ interprets R, and $\underline{0}$ is the second zero in $\omega + \omega$.

Observe that for every $\mathcal{S} \in \mathcal{K}$ there is $\varphi \in \mathcal{L}(x)$ such that $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$. For example:

(a) $(\omega, <, \{2\}) \models \forall x(Cx \leftrightarrow \exists yz(Ryz \wedge Rzx \wedge \forall w(Rwx \to w = y \vee w = z)))$

(b) $(\omega + \omega, <, \{\underline{0}\}) \models \forall x(Cx \leftrightarrow \exists y(Ryx \wedge \forall z(Rzx \to \exists w(Rzw \wedge Rwx))))$

Nonetheless, $\mathcal{K}$ is unsolvable.

Proof: Let scientist Ψ solve $\mathcal{O} = (\omega + \omega, <, \{\underline{0}\})$. We show that for some $i \in \omega$, Ψ does not solve $(\omega, <, \{i\})$. By Lemma (9) there is $\sigma \in$ SEQ, $p : var(\sigma) \to \omega + \omega$, and $\varphi_0 \in \mathcal{L}(x)$ such that:

(11) (a) $\mathcal{O} \models \bigwedge\sigma[p]$;

[5] Both results are based on an idea found in [2].

(b) $\mathcal{O} \models \forall x(Cx \leftrightarrow \varphi_0)$.
(c) for all $\gamma \in$ SEQ, if
  i. $\sigma \subseteq \gamma$ and
  ii. $\mathcal{O} \models \exists x_1 \ldots x_k \bigwedge \gamma[p]$, where $var(\gamma) - var(\sigma) = \{x_1 \ldots x_k\}$
then $\Psi(\gamma) = \varphi_0$.

It is evident that:

(12) For all but at most one $i \in \omega$, $(\omega, <, \{i\}) \not\models \forall x(Cx \leftrightarrow \varphi_0)$.

It is also easy to verify that there are infinitely many $\ell \in \omega$, complete assignments h to $(\omega, <, \{\ell\})$, and environments e for $(\omega, <, \{\ell\})$ via h such that:

(13) (a) $\sigma \subseteq e$
  (b) for all $j > length(\sigma), \mathcal{O} \models \exists x_1 \ldots x_k \bigwedge \bar{e}_j[p]$, where $var(\bar{e}_j) - var(\sigma) = \{x_1 \ldots x_k\}$

By (11)a,c and (13), $\Psi(\bar{e}_j) = \varphi_0$ for cofinitely many $j \in \omega$. So by (12), Ψ does not solve $(\omega, <, \{\ell\})$ for some choice of ℓ. ■

4 Theorems on the discovery of first-order intensions

We present four theorems on the solvability of classes of models, in the sense of the paradigm just introduced. Of particular interest are classes that arise from theories in the following way.

(14) DEFINITION:

(a) Let $T \subseteq \mathcal{L}$ be given. The class $\{\mathcal{S} \mid \mathcal{S} \models T\}$ is denoted by MOD(T).

(b) Let collection $\mathcal{K}$ of models be given. If $\mathcal{K} =$ MOD(T) for some $T \subseteq \mathcal{L}$ then $\mathcal{K}$ is called *elementary.* If $\mathcal{K} =$ MOD(T) for some recursively enumerable $T \subseteq \mathcal{L}$ then $\mathcal{K}$ is called *recursively axiomatizable.*[6]

For simplicity we limit attention to recursively axiomatizable classes; extension to arbitrary elementary classes is straightforward (see [18, 20] for analogous developments).

[6] A theorem due to Craig [6] shows that for every recursively enumerable $T \subseteq \mathcal{L}$ there is recursive $T' \subseteq \mathcal{L}$ such that T and T' have the same deductive consequences. Consequently, $\mathcal{K}$ is recursively axiomatizable iff $\mathcal{K} =$ MOD(T) for some recursive $T \subseteq \mathcal{L}$.

4.1 A universal scientist

It is not difficult to specify recursively axiomatizable classes of models that can be solved neither by computable nor by uncomputable scientist (for example, any recursively axiomatizable class containing the models of Example (10)). Consequently, no scientist is universal in the sense of solving all such classes. On the other hand, the following theorem shows that there is a mechanical, universal scientist in the weaker sense of solving all recursively axiomatizable classes that are solvable (by machine or nonmachine). To state the theorem, let Turing Machines be conceived as enumerating subsets of $\mathcal{L}$, and let T_i denote the set of sentences enumerated by the i^{th} machine.

(15) THEOREM: There is a computable function $f : \omega \times \text{SEQ} \rightarrow \mathcal{L}(x)$ such that for all $i \in \omega$, if $\text{MOD}(T_i)$ is solvable (by either computable or noncomputable scientist), then $\lambda\sigma f(i, \sigma)$ solves $\text{MOD}(T_i)$.

In the theorem, $\lambda\sigma f(i, \sigma)$ represents the computable scientist that results from parameterizing f with an index for theory T_i.

Proof: The function f is computed by a simple modification to the algorithm M presented in [20, Section 3]. Specifically, it suffices to:

(a) set $\mathbf{P}$ in M's oracle equal to the class of all sentences of form $\forall x(Cx \leftrightarrow \varphi)$ where $\varphi \in \mathcal{L}(x)$, and

(b) delete the first clause from the definition of T,$\mathbf{P}$-*potential* in the description of M's behavior (thereby allowing M to stabilize on a theory that follows logically from T_i along with the data in the current environment).

Verification of the universality of the resulting algorithm follows essentially the same proof as that given in [20, Theorem 18]. ■

Theorem (15) shows that noncomputable scientists have no advantage over their computable counterparts when it comes to solving recursively axiomatizable classes of models. This fact may be expressed as follows.

(16) COROLLARY: Let $\mathcal{K}$ be a recursively axiomatizable class of models. If $\mathcal{K}$ is solvable then some computable scientist solves $\mathcal{K}$.

Corollary (16) has the following practical consequence. Suppose that a software engineer is thinking of writing a program to solve a certain, recursively axiomatizable class $\mathcal{K}$ of models. Before proceeding, she wishes to confirm that the task is possible in principle. For this purpose it is sufficient to conceive of an arbitrary scientist (not necessary computable) that solves $\mathcal{K}$. This guarantees that a program can ultimately be found to solve $\mathcal{K}$.

4.2 Nonuniversality for nonelementary classes

The validity of Theorem (15) hinges on the elementary character of the model-classes in question. Indeed, the next theorem shows that for nonelementary classes, mechanical scientists are neither universal nor in general equivalent to nonmechanical scientists. To state

the theorem, we fix the nonlogical vocabulary of $\mathcal{L}$ to be the binary relation symbol R, the constant symbol a, the constant symbol $\underline{0}$, and the unary function symbol S, along with the distinguished predicate C. We also define a collection $\mathcal{K}_0$ of models as follows. Given $\mathbf{r} \subseteq \omega^2$, we let $p_1(\mathbf{r})$ denote the first projection of $\mathbf{r}$.

(17) DEFINITION: Choose $Z \subset \omega$ to be nonarithmetical.[7] $\mathcal{K}_0$ is the class of all models of either of the following forms (where the nonlogical vocabulary of $\mathcal{L}$ is interpreted in the order R,a,$\underline{0}$,S,C).

(a) $(\omega, \mathbf{r}, i, 0, s, p_1(\mathbf{r}))$, where $i \in Z$, s is successor, and $\mathbf{r}$ is an arbitrary subset of ω

(b) $(\omega, \mathbf{r}, i, 0, s, \mathbf{f})$, where $i \notin Z$, s is successor, $\mathbf{r}$ is an arbitrary subset of ω, and $\mathbf{f}$ is an arbitrary, finite subset of ω.

(18) THEOREM:

(a) $\mathcal{K}_0$ is solvable, but not by computable scientist.

(b) For every computable scientist Ψ there is a computable scientist Φ such that $\{\mathcal{S} \in \mathcal{K}_0 \mid \Phi \text{ solves } \mathcal{S}\} \supset \{\mathcal{S} \in \mathcal{K}_0 \mid \Psi \text{ solves } \mathcal{S}\}$

Proof: Some notations will be helpful. For the first notation, we observe that for every model in $\mathcal{K}_0$, the interpretation of $\underline{0}$ and S is 0 and successor, respectively. Consequently, for every finite $\mathbf{f} \subseteq \omega$ we may choose C-free $\phi_{\mathbf{f}} \in \mathcal{L}(x)$ such that $\phi_{\mathbf{f}}$ defines $\mathbf{f}$ in every $\mathcal{S} \in \mathcal{K}_0$. As a second notation, we use $\underline{i}$ to denote the term $S \cdots S\underline{0}$ (i occurrences of S). Finally, for $\sigma \in \text{SEQ}$, *range*(σ) denotes the set of formulas appearing in σ.

Proof of part (a). We define a (noncomputable) scientist Γ that solves $\mathcal{K}_0$. Let $\sigma \in \text{SEQ}$ be given. If *range*(σ) does not contain exactly one formula of form $a = \underline{i}$ then $\Gamma(\sigma) = (x \neq x)$. Otherwise, if σ contains one formula of form $a = \underline{i}$ then:

(a) if $i \in Z$, $\Gamma(\sigma) = \exists y Rxy$;

(b) if $i \notin Z$, $\Gamma(\sigma) = \phi_{\mathbf{f}}$, where $\mathbf{f} = \{n \in \omega \mid \sigma \text{ contains a formula of form } C\underline{n}\}$.

It is easy to verify that Γ solves $\mathcal{K}_0$.

To show that no computable scientist solves $\mathcal{K}_0$, we rely on the following definitions and lemmas. Let $\text{VAR} = \{v_i \mid i \in \omega\}$ be the variables of $\mathcal{L}$, and let $g_0 : \text{VAR} \to \omega$ be such that $g_0(v_i) = i$ for all $i \in \omega$. Given $i \in \omega$ and $\gamma \in \text{SEQ}$, γ is called "i-good" just in case there is a model $\mathcal{S}$ of form $(\omega, \mathbf{r}, i, 0, s, p_1(\mathbf{r}))$ such that $\mathcal{S} \models \bigwedge \gamma[g_0]$. The following facts are easy to prove.

(19) (a) The set $\{(i, \gamma) \mid i \in \omega \text{ and } \gamma \text{ is } i\text{-good}\}$ is recursive.

(b) For all $i \in \omega$, $\gamma \in \text{SEQ}$, and atomic formulas α of $\mathcal{L}$, if γ is i-good then either $\gamma\alpha$ or $\gamma\neg\alpha$ is i-good (where juxtaposition denotes concatentation).

[7] For a definition of nonarithmetical sets along with discussion of other technical material figuring in the proofs of this section, see [23].

(20) LEMMA: Suppose that scientist Ψ solves $\mathcal{K}_0$. Then for all $i \in Z$ there is $\sigma \in \text{SEQ}$ such that:

(a) σ is i-good;

(b) for every i-good $\gamma \in \text{SEQ}$, if $\gamma \supseteq \sigma$, then $\Psi(\gamma) = \Psi(\sigma)$.

Proof of the lemma: Suppose that scientist Ψ solves $\mathcal{K}_0$, and let $i_0 \in Z$ be given. We prove a contradiction from the hypothesis that the lemma fails for this Ψ and i_0. Falsity of the lemma implies:

(21) For all $\sigma \in \text{SEQ}$, if σ is i_0-good then for some i_0-good $\gamma \in \text{SEQ}$, $\gamma \supseteq \sigma$ and $\Psi(\gamma) \neq \Psi(\sigma)$.

We shall exhibit an environment for some model in $\mathcal{K}_0$ of form $(\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}))$ that Ψ does not solve, contradicting our choice of Ψ. The environment to be constructed will be called e, and the model in question will be called $\mathcal{S}$. e will be for $\mathcal{S}$ via g_0. We construct e in stages, the mth stage devoted to $e^m \in \text{SEQ}$. It will be the case that $e^0 \subseteq e^1 \subseteq \cdots$. We take $e = \bigcup_{m \in \omega} e^m$. $\mathcal{S}$ will be defined from e. The construction will ensure that for every $m \geq 0$, for at least m many $i < \mathit{length}(e^m)$, $\Psi(\bar{e}_i) \neq \Psi(\bar{e}_{i+1})$. Consequently, Ψ does not solve e. For the construction, let $\{\alpha_i \mid i \in \omega\}$ enumerate the atomic formulas of $\mathcal{L}$.

Stage 0: Set $e^0 = \emptyset$.

Stage m+1: Suppose that e^m has been defined, and that e^m is i_0-good. By (21) choose i_0-good $\gamma \in \text{SEQ}$ such that $\gamma \supseteq \sigma$ and $\Psi(\gamma) \neq \Psi(\sigma)$. Let $j \in \omega$ be least such that $\{\alpha_j, \neg\alpha_j\} \cap \mathit{range}(\gamma) = \emptyset$. If $\gamma\alpha_j$ is i_0-good, let $e^{m+1} = \gamma\alpha_j$; otherwise, let $e^{m+1} = \gamma\neg\alpha_j$. By (19)b, e^{m+1} is well-defined (and i_0-good).

Let $\mathcal{S} = (\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}))$, where $\mathbf{r} = \{(i,j) \in \omega^2 \mid Rij \in e\}$. The construction implies that e is for $\mathcal{S}$ via g_0, and that $\Psi(\bar{e}_i) \neq \Psi(\bar{e}_{i+1})$ for infinitely many $i \in \omega$. However, $\mathcal{S} \in \mathcal{K}_0$. ■

(22) LEMMA: Suppose that scientist Ψ solves $\mathcal{K}_0$, and let $i \notin Z$ be given. Then there is no $\sigma \in \text{SEQ}$ such that:

(a) σ is i-good;

(b) for every i-good $\gamma \in \text{SEQ}$, if $\gamma \supseteq \sigma$, then $\Psi(\gamma) = \Psi(\sigma)$.

Proof of the lemma: Suppose that scientist Ψ solves $\mathcal{K}_0$, and let $i_0 \notin Z$ be given. We prove a contradiction from the hypothesis that the lemma fails for this Ψ and i_0. Falsity of the lemma implies:

(23) There is $\sigma \in \text{SEQ}$ such that:

(a) σ is i_0-good;

(b) for every i_0-good $\gamma \in \text{SEQ}$ if $\gamma \supseteq \sigma$ then $\Psi(\gamma) = \Psi(\sigma)$.

Let σ_0 be as specified by (23). Let $\mathcal{S} = (\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}))$, where $\mathbf{r} = \{(i,j) \in \omega^2 \mid R\underline{ij} \in range(\sigma_0)\}$. $p_1(\mathbf{r})$ is finite, so (since $i_0 \notin Z$) $\mathcal{S} \in \mathcal{K}_0$. Choose $k \in \omega - p_1(\mathbf{r})$, and let $\mathcal{U} = (\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}) \cup \{k\})$. $\mathcal{U} \in \mathcal{K}_0$. Then:

(24) No $\varphi \in \mathcal{L}(x)$ defines both $p_1(\mathbf{r})$ in $\mathcal{S}$ and $p_1(\mathbf{r}) \cup \{k\}$ in $\mathcal{U}$.

Let t be an environment for $\mathcal{S}$ via g_0 such that $\sigma_0 \subseteq t$. Let u be an environment for $\mathcal{U}$ via g_0 such that $\sigma_0 \subseteq u$. It is easy to verify the following:

(25) For all $j \in \omega$, both $\bar{t}_j$ and $\bar{u}_j$ are i_0-good.

From (25) and (23) it follows that for all but finitely many $j \in \omega$, $\Psi(\bar{t}_j) = \Psi(\bar{u}_j) = \Psi(\sigma_0)$. Since $\mathcal{S}, \mathcal{U} \in \mathcal{K}_0$, (24) implies that Ψ does not solve $\{\mathcal{S}, \mathcal{U}\} \subseteq \mathcal{K}_0$, contradicting our choice of Ψ. ■

Returning to the proof of part (a) of Theorem (18), suppose for a contradiction that computable scientist Ψ solved $\mathcal{K}_0$. Then, by lemmas (20) and (22):

(26) For all $i \in \omega$, $i \in Z$ if and only if there is $\sigma \in$ SEQ such that:

(a) σ is i-good;

(b) for every i-good $\gamma \in$ SEQ, if $\gamma \supseteq \sigma$, then $\Psi(\gamma) = \Psi(\sigma)$.

However, (26), (19)a, and the computability of Ψ exhibit Z as arithmetical, contradicting our choice in Definition (17). ■

Proof of part (b). Let computable scientist Ψ be given. By part (a) of the theorem, either there is $i_0 \in Z$ such that Ψ does not solve some model of form $(\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}))$, where $\mathbf{r}$ is an arbitrary subsets of ω^2, or there is $i \notin Z$ such that Ψ does not solve some model of form $(\omega, \mathbf{r}, i_0, 0, s, \mathbf{f})$ where $\mathbf{r}$ is an arbitrary subset of ω^2, and $\mathbf{f}$ is an arbitrary, finite subset of ω. Suppose the first case, the second being parallel. Define scientist Φ as follows. For all $\sigma \in$ SEQ:

$$\Phi(\sigma) = \begin{cases} \Psi(\sigma) & \text{if } a = i_0 \notin range(\sigma); \\ \exists y Rxy & \text{otherwise.} \end{cases}$$

It is easy to verify that Φ solves all models of form $(\omega, \mathbf{r}, i_0, 0, s, p_1(\mathbf{r}))$, where $\mathbf{r}$ is an arbitrary subset of ω^2, and that $\{\mathcal{S} \in \mathcal{K}_0 \mid \Phi \text{ solves } \mathcal{S}\} \supset \{\mathcal{S} \in \mathcal{K}_0 \mid \Psi \text{ solves } \mathcal{S}\}$. ■

4.3 Weak solvability

Let environment e for model $\mathcal{S}$ be given, and suppose that scientist Ψ solves e. An external observer cannot determine with certainty whether Ψ has reached the convergent stage of its investigation of e, ceasing henceforth to change conjectures. However, evidence in favor of Ψ's convergence might be available in the form of a long succession of repeated conjectures. We now introduce a criterion of inductive success that further limits the information available to an external observer about convergence. Our definition relies on the following, standard

notation. Given a collection $\mathcal{K}$ of models, $T(\mathcal{K})$ denotes the set of sentences true in every member of $\mathcal{K}$. $T(\mathcal{K})$ may thus be conceived as representing the background theory known to a scientist about the class of possible realities. In the event that $\mathcal{K}$ is axiomatizable, say by theory T_0, then $T(\mathcal{K}) = \{\theta \in \mathcal{L} \mid T_0 \models \theta\}$.

(27) DEFINITION: Let collection $\mathcal{K}$ of models, model $\mathcal{S}$, environment e for $\mathcal{S}$, and scientist Ψ be given.

(a) Ψ *weakly solves* e just in case there is $\varphi \in \mathcal{L}(x)$ such that $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$, and $T(\mathcal{K}) \models \forall x(\Psi(\bar{e}_i) \leftrightarrow \varphi)$ for all but finitely many $i \in \omega$.

(b) Ψ *weakly solves* $\mathcal{S}$ just in case Ψ weakly solves every environment for $\mathcal{S}$.

(c) Ψ *weakly solves* $\mathcal{K}$ just in case Ψ weakly solves every $\mathcal{S} \in \mathcal{K}$. In this case, $\mathcal{K}$ is *weakly solvable.*

Thus, to weakly solve e, cofinitely many of Ψ's conjectures must be equivalent (over $T(\mathcal{K})$) to some one formula that is coextensive with C in the underlying model. Plainly, solvability implies weak solvability. For recursively axiomatizable classes, the converse also holds. This is shown by the corollary to the following theorem.

(28) THEOREM: Every weakly solvable class of models is solvable.

Proof: Suppose that scientist Ψ weakly solves class $\mathcal{K}$ of models. Given $\sigma \in \mathrm{SEQ}$, let σ^- be the result of removing the last member of σ if $\mathit{length}(\sigma) > 0$; otherwise, $\sigma^- = \sigma$. Now define scientist Φ as follows. For all $\sigma \in \mathrm{SEQ}$, $\Phi(\sigma) = \Psi(\sigma)$ if $\mathit{length}(\sigma) = 0$; otherwise:

$$\Phi(\sigma) = \begin{cases} \Phi(\sigma^-) & \text{if } T(\mathcal{K}) \models \Psi(\sigma) \leftrightarrow \Psi(\sigma^-) \\ \Psi(\sigma) & \text{if } T(\mathcal{K}) \not\models \Psi(\sigma) \leftrightarrow \Psi(\sigma^-) \end{cases}$$

It is easy to see that Φ solves $\mathcal{K}$. ■

(29) COROLLARY: Let $\mathcal{K}$ be a recursively axiomatizable class of models. If $\mathcal{K}$ is weakly solvable, then some computable scientist solves $\mathcal{K}$.

Proof: Let $\mathcal{K}$ be a recursively axiomatizable class of models, and let $i_0 \in \omega$ be such that $\mathcal{K} = \mathrm{MOD}(T_{i_0})$. Suppose that $\mathcal{K}$ is weakly solvable. Then by Theorem (28), $\mathcal{K}$ is solvable. Let f be the computable function given by Theorem (15). Then the computable scientist $\lambda\sigma f(i_0, \sigma)$ solves $\mathrm{MOD}(T_{i_0}) = \mathcal{K}$. ■

Theorem (18)a shows that Corollary (29) does not hold if $\mathcal{K}$ is allowed to be nonelementary. The next theorem strengthens this result, and in fact implies the essential content of (18)a. As a preliminary, we fix the nonlogical vocabulary of $\mathcal{L}$ for the remainder of this section to be the three constant symbols $\underline{0}, \underline{1}, a$, the two binary functions symbols $\oplus, \otimes$, and the distinguished predicate C. The term $\underline{0} + \underline{1} + \cdots + \underline{1}$ (n $\underline{1}$'s) is denoted by $\underline{n}$.

(30) THEOREM: There is a collection $\mathcal{K}$ of models such that:

(a) some computable scientist weakly solves $\mathcal{K}$;

(b) no computable scientist solves $\mathcal{K}$.

Proof: Let $\{\Psi_i \mid i \in \omega\}$ be any acceptable indexing of the computable scientists.[8] For each $i \in \omega$ we shall define a subcollection $\mathcal{K}_i$ of models. $\mathcal{K}_i$ will have either one or two members. It will be the case that Ψ_i does not solve $\mathcal{K}_i$. The needed witness $\mathcal{K}$ for the theorem will be defined as the union of the $\mathcal{K}_i$. It follows immediately that no computable scientist solves $\mathcal{K}$. Finally, a computable scientist that weakly solves $\mathcal{K}$ will be exhibited.

The following definitions and notation will facilitate the construction. Given $i \in \omega$ and $\mathbf{x} \subseteq \omega$, define $\mathcal{S}(i, \mathbf{x})$ to be the model $(\omega, 0, 1, +, \times, i, \mathbf{x})$, interpreting $\underline{0}, \underline{1}, \oplus, \otimes, a, C$ respectively. Let $g_0 : \mathrm{VAR} \to \omega$ be such that for all $i \in \omega$, $g_0(v_i) = i$, and let $\{\alpha_i \mid i \in \omega\}$ recursively enumerate the atomic formulas of $\mathcal{L}$. Given $i \in \omega$ and $\mathbf{x} \subseteq \omega$, the *canonical* environment for $\mathcal{S}(i, \mathbf{x})$ is the environment e for $\mathcal{S}(i, \mathbf{x})$ via g_0 such that for all $j \in \omega$, the jth member of e is either α_j or $\neg\alpha_j$. We note:

(31) There is a mechanical procedure that inputs $i \in \omega$ and finite $\mathbf{x} \subseteq \omega$ and outputs the canonical environment for $\mathcal{S}(i, \mathbf{x})$.

Now let $i_0 \in \omega$ be given, corresponding to scientist Ψ_{i_0}. We define $\mathcal{K}_{i_0}$ by constructing in stages a canonical environment a and a set $A \subseteq \omega$. The result of the mth stage in the construction of a and A will be denoted by a^m and A^m, respectively. If the construction proceeds through infinitely many stages, then $\mathcal{K}_{i_0} = \{\mathcal{S}(i_0, A)\}$, where $A = \{j \mid C\underline{j} \in \mathit{range}(a)\}$ (A may be infinite in this case). If the construction proceeds through only m stages, then $\mathcal{K}_{i_0} = \{\mathcal{S}(i_0, A^m), \mathcal{S}(i_0, A^m \cup \{j_m\})\}$, where $j_m \in \omega - A^m$.

——— *Construction for* i_0 ———

Stage 0: $a^0 = \emptyset$. $A^0 = \emptyset$.

Stage $m + 1$*:* Suppose that a^m and A^m have been constructed, that $A^m = \{j \mid C\underline{j} \in \mathit{range}(a^m)\}$, and that a^m is an initial segment of the canonical environment for $\mathcal{S}(i_0, A^m)$. Let $j_m \in \omega$ be least such that $j_m \notin A^m$. Let b be the canonical environment for $\mathcal{S}(i_0, A^m)$, and let c be the canonical environment for $\mathcal{S}(i_0, A^m \cup \{j_m\})$ (thus, both b and c begin with a^m). Observe that if $\Psi_{i_0}(\bar{b}_j) = \Psi_{i_0}(\bar{c}_j) = \Psi_{i_0}(a^m)$ for all $j > \mathit{length}(a^m)$, then Ψ_{i_0} fails to solve at least one of $\{\mathcal{S}(i_0, A^m), \mathcal{S}(i_0, A^m \cup \{j_m\})$. In this case the construction remains at the present stage and $\mathcal{K}_{i_0}$ is defined to be $\{\mathcal{S}(i_0, A^m), \mathcal{S}(i_0, A^m \cup \{j_m\})$. Let $q \in \omega$ be least such that $q > \mathit{length}(a^m)$ and either:

(a) $\Psi_{i_0}(\bar{b}_q) \neq \Psi_{i_0}(a^m)$ or

(b) $\Psi_{i_0}(\bar{c}_q) \neq \Psi_{i_0}(a^m)$.

In case (a), set $a^{m+1} = \bar{b}_q$; otherwise, set $a^{m+1} = \bar{c}_q$. In either case set $A^{m+1} = \{j \mid C\underline{j} \in \mathit{range}(a^{m+1})\}$.

[8] For discussion of acceptable indexings, see [15].

Observe that for every $m \in \omega$, if a^{m+1} exists, then Ψ_{i_0} changes its conjecture at least m times before reaching the end of a^{m+1}.

In case the construction completes infinitely many stages, we define environment a to be $\bigcup_{m\in\omega} a^m$. In this case, define $A = \{j \mid C\underline{j} \in \mathit{range}(a)\}$, and take $\mathcal{K}_{i_0}$ to be $\{\mathcal{S}(i_0, A)\}$. It is easy to see that in this case a is the canonical environment for $\mathcal{S}(i_0, A)$, and that $\Psi_{i_0}(\bar{a}_j) \neq \Psi_{i_0}(\bar{a}_{j+1})$ for infinitely many $j \in \omega$. So in this case Ψ_{i_0} does not solve $\mathcal{K}_{i_0}$. On the other hand, suppose that the construction completes only finitely many stages, and let b, c be the environments created during the last stage entered (say, m). Then b is the canonical environment for $\mathcal{S}(i_0, A^m)$ and c is the canonical environment for $\mathcal{S}(i_0, A^m \cup \{j_m\})$. Take $\mathcal{K}_{i_0} = \{\mathcal{S}(i_0, A^m), \mathcal{S}(i_0, A^m \cup \{j_m\})$. As noted in the construction, Ψ_{i_0} converges on b and c to the same formula, and hence fails to solve at least one of them. So in this case too, Ψ_{i_0} does not solve $\mathcal{K}_{i_0}$.

Define $\mathcal{K} = \bigcup_{i\in\omega} \mathcal{K}_i$. Then no computable scientist solves $\mathcal{K}$. It remains to exhibit computable scientist Φ that weakly solves $\mathcal{K}$.

For $i \in \omega$, let A_i be the set defined by the construction for i. This set may be infinite in case the construction completes every stage; otherwise it is finite. In view of (31) it is easy to verify the following about the sets A_i.

(32) There is a computer program P with the following property. For all input $i \in \omega$, P returns $\varphi \in \mathcal{L}(x)$ such that:

  (a) φ contains only the vocabulary $\underline{0}, \underline{1}, \oplus, \otimes$;

  (b) for every $\mathcal{S} \in \mathcal{K}$ and $n \in \omega$, $n \in A_i$ iff $\mathcal{S} \models \varphi(\underline{n})$.

Given $i \in \omega$, let φ_i be as specified in (32). By our definition of $\mathcal{K}$ we have the following fact.

(33) Let $\mathcal{N}$ be the standard model of arithmetic.[9] Then for all $\theta \in \mathcal{L}$ over the vocabulary $\underline{0}, \underline{1}, \oplus, \otimes$, $\mathcal{N} \models \theta$ iff $T(\mathcal{K}) \models \theta$.

The desired scientist Φ may now be defined. Given $\sigma \in$ SEQ, let $\Gamma(\sigma) \in \mathcal{L}(x)$ be the disjunction of $\{x = \underline{n} \mid C\underline{n} \in \mathit{range}(\sigma)\}$. For all $\sigma \in$ SEQ, $\Phi(\sigma)$ is defined to be $x \neq x$ if $\mathit{range}(\sigma)$ does not include exactly one sentence of the form $a = \underline{i}$. Otherwise, $\Phi(\sigma)$ is $\varphi_i \vee \Gamma(\sigma)$, where $(a = \underline{i}) \in \mathit{range}(\sigma)$.

By (32), Φ is computable. To verify that Φ weakly solves $\mathcal{K}$, let $i \in \omega$ be given and suppose that e is an environment for $\mathcal{S}(i, \mathbf{x}) \in \mathcal{K}$. Then for all but finitely many $j \in \omega$, $\mathcal{S}(i, \mathbf{x}) \models \forall x(Cx \leftrightarrow (\varphi_i \vee \Gamma(\sigma))$. Moreover, using (33), it is easy to verify that for all but finitely many $j, k \in \omega$, $T(\mathcal{K}) \models \forall x((\varphi_i \vee \Gamma(\bar{e}_j)) \leftrightarrow (\varphi_i \vee \Gamma(\bar{e}_k)))$. ■

[9] For background discussion of the model-theory of arithmetic, see [8, Chapter 3].

5 Concluding remarks

The foregoing paradigms and theorems suggest the potential role of contemporary logical theory in the analysis of machine learning. It is evident that research within this perspective is still in its infancy, and would profit greatly from interaction with more established theoretical traditions. New techniques from Model Theory may also be required to settle questions that emerge from the framework we have presented. One such question is formulated as follows.

(34) DEFINITION: Let collection $\mathcal{K}$ of models, model $\mathcal{S}$, environment e for $\mathcal{S}$, and scientist Ψ be given.

(a) Ψ *BC-solves* e just in case for all but finitely many $i \in \omega$ there is $\varphi \in \mathcal{L}(x)$ such that $\mathcal{S} \models \forall x(Cx \leftrightarrow \varphi)$, and $\Psi(\bar{e}_i) = \varphi$.[10]

(b) Ψ *BC-solves* $\mathcal{S}$ just in case Ψ BC-solves every environment for $\mathcal{S}$.

(c) Ψ *BC-solves* $\mathcal{K}$ just in case Ψ BC-solves every $\mathcal{S} \in \mathcal{K}$. In this case, $\mathcal{K}$ is *BC-solvable.*

(35) OPEN QUESTION:

(a) What is the relation between solvability and BC-solvability among elementary and nonelementary collections of models?

(b) Under what circumstances does BC-solvability imply BC-solvability by computable scientist?

We conclude on a speculative note. Interaction may well be desirable between a model-theoretic approach to learning, on the one hand, and issues in knowledge representation, on the other. To see what is at stake, consider a sophisticated data-base, DB. Part of the knowledge stored in DB may consist of well-confirmed statements that serve as the axioms of a class of models. To augment its knowledge, DB can wait for external assistance to augment the axiom set, or it can launch its own investigation via an automated system of scientific discovery. In the latter case, DB would be wise to reflect on the prospects for successful inquiry. What guarantee is there that DB will succeed in any, arbitrary model of its axioms, that is, in any situation consistent with what DB knows so far? If DB elects to adopt some version of the "closed world assumption," what guarantee exists that DB's empirical inquiry will succeed even in just the minimal models, those it relies on to extrapolate its data to new, plausible claims? If there is no guarantee of success, can DB at least be certain that it will not stabilize on a false theory, but rather continue endlessly to advance theories, each ultimately perceived to be inaccurate? And suppose that DB's scientific discovery routine asks for an opinion about some sentence that does not follow from the available data, but does follow from some nonmonotone rule of inference. To what extent is the reliability of the routine compromised by supplying it with information of this sort?

[10] BC stands for "behaviorally correct." See [4] for an analogous definition in the recursion-theoretic context.

Such questions, and many more like them, are crucial to the confidence that DB can place in the results of an empirical investigation that it carries out to some — always incomplete — point. So we would like to equip DB with the mathematical means necessary to determine in advance the feasibility of the empirical inquiry that it contemplates.

Answers to feasibility questions depend on the kind of axioms that DB takes as a scientific starting point — whether they involve more than monadic predicates, second-order quantification, etc. The answers depend as well on the kind of data available to DB, and the criterion of success to which DB aspires. It is not unlikely that progress along these lines would be facilitated by deploying the considerable understanding that has accumulated about logical theory over the last century. This knowledge figures promiently in theoretical studies of knowledge representation.[11] Perhaps it can be deployed, as well, in learning, and used as bridge between the two disciplines.

References

[1] Angluin, D. "Learning Propositional Horn Sentences with Hints," Yale University Technical Report, 1987.

[2] Blum, L. & Blum, M. "Toward a mathematical theory of inductive inference," *Information & Control*, 28:125-155.

[3] Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. "Learnability and the Vapnik-Chervonenkis Dimension," *Journal of the ACM,* 36:4, 1989.

[4] Case, J. & Lynes, C., "Machine inductive inference and language identification," *Proceedings of the 9th Colloquium on Automata, Languages, and Programming,* Springer-Verlag, Lecture Notes in Computer Science, 140, 1982, pp. 107-115.

[5] Chang, C. C., & Keisler, H. J., *Model Theory*, North Holland, 1973.

[6] Craig, W., "On Axiomatizability Within a System," *Journal of Symbolic Logic*, 18:30-32.

[7] Ehrenfeucht, A., Haussler, D., Kearns, M. & Valiant, L., "A General Lower Bound on the Number of Examples Needed for Learning," *Information and Computation* **82** (1989) 247–261.

[8] Enderton, H., *A Mathematical Introduction to Logic*, Academic Press, 1972.

[9] Fulk, M. & Case, J. (Eds.) *Proceedings of the Third Annual Workshop on Computational Learning Theory,* Morgan-Kaufmann, 1990.

[10] Gold, E. M., "Language Identification in the limit," *Information & Computation* **10** (1967) 447-474.

[11] see, for example, [21].

[11] Haussler, D. "Learning conjunctive concepts in structural domains," *Machine Learning* 4:7-40, 1989.

[12] Haussler, D. & Pitt, L. (eds.) *COLT 88: Proceedings of the 1988 Workshop on Computational Learning Theory* (1988) Morgan-Kaufmann.

[13] Kearns, M., Li, M., Pitt, L., & Valiant, L., "On the Learnability of Boolean Formulae," *Communications of the ACM* 1987.

[14] Laird, P. "Inductive Inference by Refinement," in *Proc. of AAAI-86*, pp.472-476 AAAI, 1986.

[15] Machtey, M. & Young, P., *An Introduction to the General Theory of Algorithms*, North Holland, 1978.

[16] Michalski, R. & Stepp, R., "Learning from observation: Conceptual clustering," in R. Michalsky, J. Carbonell & T. Mitchell (Eds.) *Machine Learning: An Artificial Intelligence Approach*, Tioga, 367-404.

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

[18] Osherson, D., Stob, M. & Weinstein, S. "A universal inductive inference machine," *Journal of Symbolic Logic*, in press.

[19] Osherson, D., Stob, M. & Weinstein, S. "New Directions in Automated Scientific Discovery," *Information Sciences,*, in press.

[20] Osherson, D., Stob, M. & Weinstein, S. "A Mechanical Method of Successful Inquiry," in [9].

[21] Parikh, R. (Ed.) *Theoretical Aspects of Reasoning about Knowledge*, Morgan-Kaufmann, 1990.

[22] Rivest, R., Haussler, D. & Warmuth, M. (eds.) *COLT 89: Proceedings of the Second Annual Workshop on Computational Learning Theory* (1989) Morgan-Kaufmann.

[23] Rogers, H. *Theory of Recursive Functions and Effective Computability.* New York: McGraw-Hill, 1967.

[24] Shapiro, E. "An algorithm that infers theories from facts," *Proceedings of the Seventh International Joint Conference on Artificial Intelligence, IJCAI*, 1981.

[25] E. Shapiro, *Algorithmic Program Debugging,* MIT Press, 1983.

[26] Valiant, L. "A Theory of the Learnable," *Communications of the ACM,* 27:1134-1142, 1984.