HAIM GAIFMAN, DANIEL N. OSHERSON and SCOTT WEINSTEIN

A REASON FOR THEORETICAL TERMS*

ABSTRACT. The presence of nonobservational vocabulary is shown to be necessary for wide application of a conservative principle of theory revision.

1. INTRODUCTION

The best reason for theoretical terms, we believe, is the one defended by Putnam (1965), namely, that they allow reference to small or otherwise elusive entities that are legitimate objects of scientific study. Reasons of a more structural character are also worthy of attention, however, in the hope of shedding light on theory construction and test. Thus, it might be thought that additional predicates are sometimes necessary for the recursive axiomatization of scientific theory. It is well known that Craig (1953) refuted this conjecture by showing that if a theory has a recursively axiomatizable conservative extension then it is recursively axiomatizable.

Recursive axiomatizability, however, may be too crude a measure for detecting the utility of theoretical terms in scientific theories. Considerations of computational complexity – a vigorous research area in computer science – might provide a finer scale of theoretical utility than the recursive–nonrecursive dichotomy, and throw light on the role of theoretical terms in theory formation and test. This is not the orientation of the present paper, however.

Two different structural reasons for theoretical terms were adduced by Hempel (1963). On the one hand, Hempel noted that additional vocabulary may open the way to finite axiomatization of scientific theory. On the other hand, he suggested that new vocabulary might help to rationalize confirmation relations obtaining among sentences in the original language.

Following up Hempel's discussion, the present paper advances the following reason for theoretical terms. Theory revision in the face of disconfirming data is sometimes facilitated by theoretical vocabulary. Intuitively, it is easy to see why this might be the case. The predictions of a theory sometimes segregate into clusters that rest on distinct,

Erkenntnis **32**: 149–159, 1990.
© 1990 *Kluwer Academic Publishers. Printed in the Netherlands.*

hypothetical mechanisms. Predictive failures in such cases tend to infect well-defined subsystems of axioms in the falsified theory, sparing others. The impact of disconfirming data may thus be canalized by favoring theoretical revisions that respect intact fragments of the discredited theory. It is often the presence of theoretical vocabulary that allows coherent subsystems of axioms to emerge in theory development, making it easier to apportion praise and blame among axioms in the light of experience.

The preservation of axioms that have played little or no role in false predictions is a conservative strategy of theoretical revision. Reliance on such a strategy provides no guarantee of scientific success since it may prevent overhaul of a theory that is fundamentally mistaken. Nonetheless, conservative theory revision is a natural policy inasmuch as it provides a reasonable guide to the choice of axioms for modification or suppression. Consequently, the use of theoretical vocabulary can be partly vindicated by showing that it is sometimes necessary for the successful application of a conservative policy.

In the present paper we show the mathematical feasibility of this kind of vindication. This is achieved by formally defining a model of scientific inquiry in which conservative theory revision is possible, but only at the cost of introducing nonobservational terms into the theory. Once the model is defined, the necessity of nonobservational terms for conservative theory revision becomes a mathematical necessity. It cannot be proved, of course, that our model is a faithful and appealing representation of scientific inquiry. But at least we will have shown that – in contrast to the approach based on recursive axiomatizability, which fails for purely mathematical reasons – the vindication of theoretical terms via conservative practice is feasible.

We shall in fact focus on a strong form of conservative theory revision. In special cases coherent fragments of a larger theory may be finitely axiomatized and thus reduced, via conjunction, to a single sentence. Preservation of such a fragment amounts to retention in the successor theory of the corresponding axiom. In these circumstances, conservative theory revision takes a particularly simple form, summarized by the following injunction.

> *Strong conservative policy*: Preserve axioms from a falsified theory that are not individually contradicted by the available data.

A formal model of scientific inquiry will be exhibited in which this strong conservative policy can be successfully maintained, but only if theories include nonobservational vocabulary.

It should be observed that strong conservative policy, by itself, imposes little constraint on theory revision. For, by making use of theoretical vocabulary virtually any scientific theory can be reduced to a single axiom (see Craig and Vaught, 1958); and strong conservatism allows single-axiom theories to be totally revised in the face of disconfirming data. In contrast, in the absence of theoretical vocabulary, many scientific theories can be axiomatized only non-finitely; and strong conservatism will often reduce the options for revising such theories when disconfirming data arise (since no individual axiom need be contradicted by data that contradict an entire theory). Indeed, we shall see that in certain cases adherence to strong conservative policy in the absence of theoretical vocabulary so constrains theory revision as to render scientific success impossible.

Our discussion proceeds as follows. Section 2 introduces the framework within which our analysis is carried out. Section 3 formulates within the given framework our claim about conservatism and theoretical terms. Proof of the claim comes in Section 4. Concluding remarks occupy Section 5.

2. STRUCTURE IDENTIFICATION

2.1. *Overview*

In what follows we use a first-order framework in order to represent a process in which established theories are confronted (and possibly revised) by newly discovered facts. Terms such as "environment" or "scientist" will be employed to refer to technical constructs in the framework which represent, or mimic, the entities in question. Thus, within our model an environment is a sequence of basic facts, arranged in the temporal order of their discovery. Similarly, a scientist is represented by any system that, given a sequence of observed facts, chooses a theory; hence scientists are identified in our model with functions from initial chunks of environments into theories (where a theory is just a set of sentences in some fixed, presupposed language). We do not intend these definitions as explications of the concepts in question, nor do we pretend to have modeled all important aspects of

scientific inquiry (for example, we have ignored the role of scientists in experiment planning, among many other topics). We have focussed only on those aspects of scientific inquiry which are relevant to our conception of conservative theory revision.

As a preliminary we fix a first-order language L containing an equality sign, $=$, and whose non-logical vocabulary consists of countably many predicates, function symbols and individual constants. The individual variables of L are $x, y, z, v_0, v_1, \ldots$. By a "sublanguage of L" we mean any first-order language L', with equality, whose vocabulary is included in that of L. Terms and formulas are defined as usual. A basic formula is understood to be an atomic formula or its negation. We rely as well on the following conventions.

> By "assignment" is meant a mapping of the variables of L into the domain of whatever structure is in question.
>
> Let L' be a sublanguage of L (possibly L itself). By "structure for L'" is meant a countable structure that interprets the nonlogical constants of L' (and no other ones). L' is said to be "the language of" a structure S just in case S is for L'.

Uncountable structures do not enter the discussion. For conceptual difficulties that arise in extending the present framework to the uncountable case, see Osherson and Weinstein (1986a, Section 6.1).

2.2. *Environments*

DEFINTION 2.2A: (i) An *environment* is an omega-sequence of basic L-formulas.

(ii) The set of (basic) formulas occurring in an environment e is denoted: set(e).

(iii) Let structure S interpret L'. An environment e is said to be *for* S just in case there is an assignment h onto $|S|$ such that set(e) $= \{B: B$ is a basic L'-formula and $S \models B[h]\}$.

On our conception of environment, variables and names play the role of observed objects, and no properties of objects beyond those expressed by basic formulas of L' are available to the scientist. The existential sentences witnessed by these latter formulas may be construed as the observation sentences engendered by the underlying

structure S. Individuation of objects may proceed by recording the equalities and inequalities that appear in a given environment.

An environment for a structure S provides virtually complete information about S. This is the content of the following lemma; its proof is given in Osherson and Weinstein (1986a, Lemma 3.1A).

LEMMA 2.2A: If S and R are two structures and e is an environment which is for both S and R, then S and R are isomorphic.

2.3. *Scientists*

Scientists are construed as dispositions to convert finite initial segments of a given environment into hypotheses about the structure giving rise to that environment. Hypotheses take the form of axiom-sets drawn from L.

DEFINITION 2.3A: (i) The set of all finite sequences of basic L-formulas is denoted: SEQ.

(ii) The set of basic L-formulas appearing in $\sigma \in SEQ$ is denoted: $\text{set}(\sigma)$.

(iii) A (*formal*) *scientist* is any total function from SEQ to the power set of all L-sentences.

Given $\sigma \in SEQ$ we let "$\&\sigma$" denote the conjunction of $\text{set}(\sigma)$, in order of appearance in σ.

2.4. *Axiomatization of a structure*

We now formulate the distinction between adequate theories that employ theoretical vocabulary and those that do not. For a given structure S, let $\text{Th}(S)$ be the set of all sentences (in the language of S) which are true in S. Let languages L' and L'' with $L' \subseteq L'' \subseteq L$ and structure S for L' be given. By an "expansion" of S to L'' is meant any structure R such that

1. R is a structure for L'';
2. $\text{domain}(R) = \text{domain}(S)$;
3. R makes the same assignments as S to L'.

DEFINITION 2.4A: Let $L' \subseteq L'' \subseteq L$ be given. Let S be a structure for L', and let T be a set of L''-sentences.

(i) *T axiomatizes S* just in case:

(a) $\mathrm{Th}(S) = \{A: A \text{ is a sentence of } L' \text{ and } T \models A\}$.

(b) There exists an expansion of S to L'' which satisfies T.

(ii) *T axiomatizes S directly* just in case:

(a) T axiomatizes S; and

(b) $T \subseteq L'$.

Thus, to axiomatize S, the set of L'-sentences deducible from T must coincide with those made true by S. Moreover, the supplementary vocabulary of T must be interpretable in the domain of S in such a way that every sentence of T is satisfied in the expanded structure. It is not required that the consequences of T coincide with the theory of the expanded structure. That is, theoretical terms may be only "partially interpreted" by the theory in which they occur.

2.5. *Criterion of scientific success*

The success criterion of the present paradigm is known as "identification". We formulate two versions of this concept corresponding to direct and indirect axiomatization of structures. A preliminary definition is needed. Let the set of natural numbers be denoted: N.

DEFINITION 2.5A: Let scientist Φ and environment e be given. Let T be a set of L-sentences.

(i) The initial finite segment of length $n \in N$ in e is denoted: $e \mid n$.

(ii) *Φ converges on e to T* just in case $\Phi(e \mid n) = T$ for all but finitely many $n \in N$.

DEFINITION 2.5B: Let scientist Φ and structure S be given.

(i) *Φ identifies S* [*directly*] just in case for each environment e for S there is a set T of L-sentences such that:

(a) Φ converges on e to T; and

(b) T axiomatizes S [directly].

(ii) Let K be a collection of structures. Φ *identifies* K [*directly*] just in case Φ identifies every $S \in K$ [directly].

(iii) A collection K of structures is *identifiable* [*directly*] just in case some scientist identifies K [directly].

In the definition, K may be conceived as a class of theoretical possibilities.

Examples of identifiable and unidentifiable collections of structures are provided in Osherson and Weinstein (1986a). For a spectrum of alternative identification criteria, see Osherson and Weinstein (1986b).

The difference between direct and indirect identification is that in the former the scientist chooses a theory in the language L' of the given structures, while in the latter the chosen theory can be in a richer language. Any scientist Φ who identifies the structures indirectly can be changed into a scientist Θ who identifies them directly, namely: $\Theta(e \mid n)$ consists of all sentences in L' which are logical consequences of $\Phi(e \mid n)$. Hence, a collection of structures is identifiable directly if and only if it is identifiable *tout court*. The restrictive character of direct identification emerges only in the context of theory revision. To this we now turn.

3. THEORY REVISION

3.1. *Conservative identification*

The next definition embodies the strong conservative policy discussed in Section 1.

DEFINITION 3.1A: Let scientist Φ and structure S be given.

(i) Φ *identifies* S *conservatively* just in case:

(a) Φ identifies S; and

(b) For all environments e for S, for all $n \in N$, and for all L-sentences $s \in \Phi(e \mid n)$, s is not in $\Phi(e \mid n+1)$ if and only if s logically contradicts $\& e \mid n+1$. (In other words, any previously accepted hypothesis is omitted if and only if it contradicts, by itself, the accumulated data.)

(ii) Let K be a collection of structures. Φ *identifies* K *conservatively* just in case Φ identifies conservatively every $S \in K$.

Observe that conservative identification imposes no conditions on the formation of new axioms.

3.2. *Conservatism and direct identification*

Scientists who eschew theoretical terms will attempt direct identification of the structures underlying their environments. Such a strategy, however, is sometimes incompatible with conservative theory revision, in the sense of Definition 3.1A. This is the content of the following proposition, which formulates the promised reason for theoretical terms.

PROPOSITION: There is a collection K of structures such that:

(a) K is identifiable conservatively; but

(b) any scientist that identifies K directly fails to identify K conservatively.

Moreover, K can be chosen so that each of its structures interprets the same, finite vocabulary.

Thus, there are conservatively identifiable collections of structures that are not identifiable both conservatively and directly.

4. PROOF OF THE PROPOSITION

We first specify the collection K that witnesses the proposition, and then verify that K satisfies clauses (a) and (b).

K consists of countably many structures $\{S_{\infty,\infty}\} \cup \{S_{m,n}\colon\ m, n \geq 1\}$. The vocabulary of all the structures is $=$ and R, where R is a binary relation symbol. Let $Z = \{\ldots -2, -1, 0, 1, 2, \ldots\}$, and let $M = \{a_0, a_1, \ldots\}$ be a fixed, countable set disjoint from Z. The domain of $S_{\infty,\infty} = Z \cup M$. The interpretation of R in $S_{\infty,\infty}$ is $\{\langle p, p+1\rangle\colon\ p \in Z\}$. The domain of $S_{m,n} = \{0, \ldots, m-1\} \cup \{a_0, \ldots, a_{n-1}\}$. The interpretation of R in $S_{m,n}$ is $\{\langle p, p+1\rangle\colon p < m-2\} \cup \{\langle m-1, 0\rangle\}$. Thus, $S_{m,n}$ consists of an R-cycle of length m, along with n additional points.

Let $C(x)$ be the wff $(\exists y)(Rxy)$. Then, in every model in K, $C(x)$ defines the members belonging to the cycle, or to Z. Call such members *connected*.

The following properties hold.

(I) For any Θ in the language of K, if $S_{\infty,\infty} \models \Theta$, then there exists a natural number k such that $S_{m,n} \models \Theta$ for all $m, n \geq k$.

Property (I) is derivable either via Ehrenfeucht–Fraisse games (one can show that we can choose $k = 2^d$, where d is the number of quantifiers in Θ) or directly via Gaifman's (1982) locality characterization of first order properties. ($S_{\infty,\infty}$ and $S_{m,n}$ are locally the same when m and n are sufficiently large with respect to the neighborhoods.)

(II) If e is an environment for $S_{m,d}$ then $\text{set}(e) \cup \text{Th}(S_{m,n})$ is consistent for every $n \geq d$.

This follows from the fact that for every k, $e \mid k$ is extendible to an environment for $S_{m,n}$. It is necessary only to add the $n - d$ "missing" members.

LEMMA 1: No conservative scientist identifies K directly.

Proof: Assume that Φ is a conservative scientist that identifies $S_{\infty,\infty}$ directly. We shall show that Φ fails to identify $S_{m,n}$ for some $m, n \geq 1$. Let e be an environment for $S_{\infty,\infty}$. Then for some k, $\Phi(e \mid k)$ axiomatizes $S_{\infty,\infty}$ directly.

Let $v = v_0, \ldots, v_{d-1}$ be the string of all variables occurring in $e \mid k$. Then there is a finite conjunction Θ of members of $\Phi(e \mid k)$ such that:

$$\Theta \models (\exists^{>d} w)(\sim C(w)),$$

where $\exists^{>d}$ means "there exists more than d". Then $S_{\infty,\infty} \models \Theta \;\&\; (\exists v)(\&\, e \mid k)$. Hence by (I) there exists $S_{m,n}$ with $m, n > d$ such that:

$$S_{m,n} \models \Theta \;\&\; (\exists v)(\&\, e \mid k).$$

Extend $e \mid k$ to an environment e' for $S_{m,d}$ by adding if necessary connected members, completing the cycle, and adding if necessary nonconnected members. By (II), e' is consistent with $\text{Th}(S_{m,n})$ and hence with Θ. So by conservatism, for all $r > k$, $\Phi(e' \mid r)$ contains all the conjuncts of Θ. But then for all $r > k$, $\Phi(e' \mid r) \models (\exists^{>d} w)(\sim C(w))$. Since $S_{m,d} \models \sim(\exists^{>d} w)(\sim C(w))$, Φ fails to identify $S_{m,d}$. ■

LEMMA 2: Any class of finite models over a finite vocabulary is identifiable directly and conservatively.

Proof: Define scientist Φ as follows. For all $\sigma \in SEQ$, if length$(\sigma) = 0$, then $\Phi(\sigma) = \emptyset$. Otherwise, if γ extends σ by one basic formula, then $\Phi(\gamma)$ is obtained from $\Phi(\sigma)$ by deleting all wffs which are inconsistent with γ and adding the sentence:

$$(\exists u)(\& \gamma \& (\forall x)(x = u_0 \vee \cdots \vee x = u_{k-1}))$$

where $u = u_0, \ldots, u_{k-1}$ is the string of all variables occurring in γ, if this sentence is not already a consequence of the sentences remaining in $\Phi(\sigma)$ after the above deletion is made. It is clear that Φ conservatively identifies any class of finite models. ■

LEMMA 3: K is identifiable conservatively, indirectly.

Proof: Add to the vocabulary of K a new binary relation symbol T. Let Θ be the sentence asserting that T is a transitive irreflexive relation over the connected members which extends R. Formally, $\Theta = \Theta_1 \& \Theta_2 \& \Theta_3 \& \Theta_4$, where:

$$\Theta_1 = (\forall x, y)(T(x, y) \rightarrow C(x) \& C(y)),$$
$$\Theta_2 = (\forall x, y, z)(T(x, y) \& T(y, z) \rightarrow T(x, z)),$$
$$\Theta_3 = (\forall x)(\sim T(x, x)), \text{ and}$$
$$\Theta_4 = (\forall x, y)(R(x, y) \rightarrow T(x, y)).$$

Define Ω as follows. For all $\sigma \in SEQ$, $\Omega(\sigma) = \{\Theta \& \Gamma : \Gamma \in \mathrm{Th}(S_{\infty,\infty})\}$ if σ does not contradict Θ; else, $\Omega(\sigma) = \Phi(\sigma)$ where Φ is the scientist provided by Lemma 2.

If e is an environment for $S_{\infty,\infty}$, then Θ will never be contradicted (for, we can interpret T as the less-than relation over Z), and hence Ω will identify $S_{\infty,\infty}$. On the other hand, if e is an environment for some $S_{m,n}$, then for some i, $e \mid i$ implies the existence of a cycle. This contradicts Θ because the transitivity of T will imply $T(x, x)$. Hence all members of $\{\Theta \& \Gamma : \Gamma \in \mathrm{Th}(S_{\infty,\infty})\}$ may be dropped without violation of conservatism. Ω then switches to simulating Φ of Lemma 2, which conservatively identifies $\{S_{m,n} : m, n \in N\}$. ■

5. CONCLUDING REMARKS

Conservatism is only one among several policies that plausibly guide theory formation and revision. Others are discussed in Osherson and Weinstein (1986a, Section 5; 1989, Section 5). We suspect that the role of nonobservational vocabulary in exploiting these policies is

fundamental to understanding the impact of theoretical terms on hypothesis formation and test.

It seems clear that investigation of such matters is best pursued in the context of precise models of scientific inquiry. The present results are embedded in a particular paradigm of this nature, but a vast range of alternatives may be defined. Analysis and comparison of these paradigms may help to clarify a variety of epistemological issues.

NOTE

* Support for this research was provided by the Office of Naval Research under contract No. N00014-87-K-0401. We thank Daniel Andler and Clark Glymour for helpful discussion.

REFERENCES

Craig, W.: 1953, 'On Axiomatizability Within a System', *Journal of Symbolic Logic* **18**, 30–32.

Craig, W. and R. S. Vaught: 1958, 'Finite Axiomatizability Using Additional Predicates', *Journal of Symbolic Logic* **23**, 289–308.

Gaifman, H.: 1982, 'On Local and Nonlocal Properties', in Stern (ed.), *Logic Colloquium*, North Holland, pp. 105–132.

Hempel, C. G.: 1963, 'Implications of Carnap's Work for the Philosophy of Science', in P. A. Schlipp (ed.), *The Philosophy of Rudolf Carnap*, Open Court, pp. 685–707.

Osherson, D. and S. Weinstein: 1986a, 'Identification in the Limit of First Order Structures', *Journal of Philosophical Logic* **15**, 55–81.

Osherson, D. and S. Weinstein: 1986b, 'Identifiable Collections of Countable Structures', *Philosophy of Science*, in press.

Osherson, D. and S. Weinstein: 1989, 'Paradigms of Truth Detection', *Journal of Philosophical Logic* **18**, 1–42.

Putnam, H.: 1965, 'Craig's Theorem', *The Journal of Philosophy*, **LXII**, 251–260.

Manuscript submitted March 25, 1987
Final version received June 14, 1988

E 10-237
MIT
Cambridge, Mass. 02139
U.S.A.