Brit. J. Phil. Sci. **18** (1967), 223–229 *Printed in Great Britain*

The Implicit Definition of Theoretical Terms*

by JOHN A. WINNIE

I

It is nowadays almost a truism that a theory which merits the title 'scientific' must be either confirmable or falsifiable by reports of possible observations. Above all, it is the theory's observable consequences that either justify its acceptance or reveal its inadequacies. Yet those scientific theories which have best demonstrated their fruitfulness are also those containing terms (which I shall call 'theoretical terms') which purportedly designate entities which are not observable, and, in some cases it seems, unobservable in principle. Here we have the basis of an epistemological problem which has been of perennial concern to philosophers of science: how do these terms which designate unobservables come to do so?[1] One can convey to others the sense of 'green' or 'cold' by, roughly, putting them in a position to see a patch of grass or touch a piece of ice. Such or similar methods are not available, however, for teaching the sense of terms like 'electron' or 'neutron'. To be sure, the presence of electrons can be indicated by a variety of observable happenings, such as white patches on the screen of a cathode ray tube; nevertheless, this is quite a different matter from observing an electron. But, even granting this, some philosophers believe that the above example does furnish a clue as to how we do come to understand what theoretical terms designate. For as one gradually learns which situations indicate the presence of, say electrons, the term 'electron' comes to be more and more meaningful; we learn that 'electron' designates a class of physical entities which, under certain observable conditions, illuminate screens of cathode ray tubes, produce white traces in cloud chambers, etc.; we learn that 'electron' designates *the* things that behave in just these particular ways under such-and-such conditions. So although we still do not have knowledge of the designation of 'electron' which is as direct and complete as our knowledge of the designations of 'green' or 'cold', yet we have an indirect and partial interpretation for the term which at least informs us of *some* of the characteristics of its designata.

* Received 27.x.66

[1] There are, of course, some philosophers who deny that theoretical terms *do* designate theoretical entities, or anything at all. But my purpose here is to examine problems posed for the realist, not to adjudicate the realism-instrumentalism issue.

It follows, moreover, from this account that theoretical terms have no significance apart from the theory in which they are imbedded. For it is the *theory*, after all, that informs us that it is *electrons* (not tiny pebbles) that reveal their presence in cathode ray tubes and cloud chambers. It is the lawlike connections which the theory asserts to hold between electrons, other theoretical entities, and observables which provides the basis for the partial interpretation of the term 'electron' (and the theory's other theoretical terms as well). This is sometimes put succinctly by saying that the theoretical terms are *implicitly* defined by the nomological network (the system of lawlike statements) in which they are imbedded. The theoretical terms acquire their empirical content by 'upward seepage' or 'osmosis' (Herbert Feigl's phrases) from the observation terms (cf. [3], pp. 34-36; [5], p. 87; and [6], pp. 46-52). Before examining this view in more detail, let us state matters somewhat more precisely.

Following Carnap [1], let the primitive terms of the theory be divided into two classes: the theoretical terms and the observation terms.[1] Let the language in which the theory is formulated be first-order, and the theory be given by a finite set of axioms. The interpretation of the primitive observation terms is assumed completely specified, perhaps given ostensively, or in some other fairly 'direct' way. Now the axioms of the theory can be said to preclude certain interpretations for the theoretical terms, namely those which make the theory false. Thus, the axioms restrict the set of all possible interpretations of theoretical terms to a much narrower set—the set of true interpretations—and this, I think, is a fairly accurate explication of what is sometimes meant by saying that the axioms 'implicitly' define these terms (certainly by Nagel [5], p. 87 and Pap [6], pp. 51-52).

Were it not for the assumption that the observation terms have a fixed empirical interpretation, this would be a patently inadequate account of the empirical significance of theoretical terms. For since we have assumed our theory couched in the language of quantification theory, Lowenheim's theorem assures us of the existence of a true *arithmetical* interpretation of the theory if the theory be satisfiable at all. Once we assume a fixed empirical interpretation for the observation predicates, however, Lowenheim's result no longer applies. But, even though Lowenheim's theorem cannot be used in this instance, can we arrive at numerical interpretations of theoretical terms via another route? The following considerations would seem to indicate a negative answer to this question. First, any adequate reconstruction of a physical theory will contain sentences (called by Carnap and others, 'correspondence postulates') which contain both theoretical

[1] Later, for the sake of greater generality, three classes of predicates will be distinguished.

and observation terms. Any admissible interpretation of the theory's theoretical terms, then, must be such as to make these sentences come out true while retaining a fixed interpretation for the observation terms. But then the correspondence rules would truly assert connections between classes of theoretical entities and classes of observational entities—seemingly an impossible situation if theoretical entities are taken to be numbers. Furthermore, the theory may contain relation symbols which, under their intended interpretations, again relate theoretical and observational entities. So we would seem to have some assurance that true numerical interpretations for theoretical terms are not generally forthcoming. The theory's correspondence rules, in particular, would seem to weave an empirical web about the theoretical terms, guaranteeing that they designate classes of physical entities—not numbers. But if the result to be presented here is correct, then, under certain trivial assumptions, a true numerical interpretation for the theoretical terms will always be forthcoming (if there exists a true physical interpretation). Furthermore, the interpretation of the observation terms is not affected as the theoretical terms are now construed as designating classes of numbers. In short, the fact that an empirical interpretation of the observation terms is taken as fixed, does not rule out true numerical interpretations for the theoretical terms. As far as the doctrine of implicit definition goes, theoretical entities might just as well be construed as numbers. In the next section, this and a related result will be stated with proofs sketched.

II

Let the language $\mathscr{L}$ in which our theory $\mathscr{T}$ is formulated be first-order and contain, for simplicity's sake, a finite number of predicates as the only descriptive signs. These predicates are divided into three classes: $P_0 = \{O_1, O_2, \ldots O_j\}$, $P_t = \{T_1, T_2, \ldots T_k\}$, and $P_m = \{M_1, M_2, \ldots M_l\}$. As will become clear later, these are the observation terms, theoretical terms, and 'mixed' terms, respectively. Rather than formulating $\mathscr{T}$ in a two-sorted language, we shall also introduce the one-place predicates '*Th*' and '*Ob*' which, intuitively, are read as 'is a theoretical entity' and 'is an observational entity', respectively. A *possible model* for $\mathscr{L}$ will then by any sequence $\langle \mathrm{Th} \cup \mathrm{Ob}, \mathrm{Th}, \mathrm{Ob}, \mathrm{O}_1, \ldots, \mathrm{O}_j; \mathrm{T}_1, \ldots, \mathrm{T}_k, \mathrm{M}_1, \ldots, \mathrm{M}_l \rangle$, where $\mathrm{Th} \cup \mathrm{Ob}$ is a non-empty class, $\mathrm{Th} \cap \mathrm{Ob}$ is empty, the O_i's are either classes or classes of n-tuples of members of Ob, the T_i's are either classes or classes of n-tuples of members of Th, and the M_i's are either classes or classes of n-tuples of $\mathrm{Th} \cup \mathrm{Ob}$. If O_i is a one-place predicate, then O_i is a class; if O_i is an n-place predicate then O_i is a class of n-tuples; similarly for the

T_i's and M_i's. Th is assigned to *Th* and Ob to *Ob*. The notion of the *truth* of a sentence S in $\mathscr{L}$ with respect to a possible model is defined in the usual way. A possible model is a *model* of a set of sentences of $\mathscr{L}$ if and only if every sentence in the set is true in $\mathscr{L}$ with respect to that possible model. We now have the following as a theorem:

Th. 1. Let $A = \langle Th_1 \cup Ob_1, Th_1, Ob_1, O_1, \ldots, O_j, T_1, \ldots, T_k, M_1, \ldots, M_1 \rangle$ be a model of $\mathscr{T}$ such that for some non-empty T_i, some element of Th_1 is not in T_i if T_i is a class or not a component of a n-tuple in T_i if T_i is a class of n-tuples. Then there is a relational system $A^* = \langle Th_1 \cup Ob_1, Th_1, Ob_1, O_1, \ldots, O_j, T_1^*, \ldots, T_k^*, M_1^*, \ldots, M_1^* \rangle$ which is also a model of $\mathscr{T}$ and is such that $T_i \neq T_i^*$.

Proof: Let T_i be the class which does not contain some element, say q, of Th_1. Let r be an element of T_i (or a component of some n-tuple in T_i) which is distinct from q. Let ϕ be a one-one function of Th_1 onto itself such that $\phi(r) = q$. We define the function ψ as follows. If $x \in Th_1$, then $\psi(x) = \phi(x)$; if $x \in Ob$, then $\psi(x) = x$. A* is now formed by replacing the elements of the classes (or components of classes of n-tuples) in the sequence A by their images under ψ. Since ψ is a one-one mapping of $Th_1 \cup Ob_1$ onto itself, A and A* are isomorphic. Hence, since A is a model of $\mathscr{T}$, A* is a model of $\mathscr{T}$. Furthermore, as can be seen from the construction, $T_i \neq T^*_i$.

The effect of *Th. 1* is to demonstrate the existence of alternate *physical* models for any theory $\mathscr{T}$ which is reconstructed along the lines given here. For if Ob_1 and Th_1 are taken as classes of observational and theoretical entities, then A* will be an alternate physical model of $\mathscr{T}$. Notice that in the construction of A* the interpretation of the observation terms is not altered, and the observational relata of the 'mixed' classes (the M_i's) remain as before. But at least one theoertical term (and for most theories, all of the theoretical terms) has received a different interpretation. From this result it follows that no amount of adding correspondence rules to a theory can ever result in 'pinning down' a unique physical interpretation for the theoretical terms of the theory.[1] So if a theory involving sub-microscopic particles, say, were reconstructed along the lines given above, the axioms (however numerous) would never specify *the* class of electrons, or *the* class of neutrons, for example. There would always be, given one true interpretation of the theory, another true interpretation in which the term 'electron' would now apply to some entities which were previously called 'neutrons'.[2]

[1] The addition of postulates may, of course result in the explicit definability of the theoretical predicates in terms of observation predicates, but this would result in violating the condition that the theoretical and observational entities are distinct.

[2] The construction would proceed along the lines of the proof of *Th.1*. We would assume that, under the initial model for the theory, 'electron' is not assigned to the class of all of the theoretical entities, and that the classes assigned to 'electron' and 'neutron' are

But as far as Th.1 goes, the theoretical terms still retain a physical interpretation. The next theorem assures us of the existence of a true *numerical* interpretation for the theoretical terms, if there is a true physical interpretation of $\mathscr{T}$.

Th. 2. Let $A = \langle Th_1 \cup Ob_1, Th_1, Ob_1, O_1, \ldots, O_j; T_1, \ldots, T_k; M_1, \ldots, M_1 \rangle$ be a model of $\mathscr{T}$ such that Th_1 is non-empty. Then there is a relational system.

$A^* = \langle Ar_1 \cup Ob_1, Ar_1, Ob_1, O_1, \ldots, O_j; T_1^*, \ldots, T_k^*; M_1^*, \ldots, M_1^* \rangle$ such that A* is also a model of $\mathscr{T}$ where Ar_1 is a class of arithmetical entities.

Proof: Let Ar_1 be a class of numbers of the same cardinality as Th_1 (by the numeration theorem, Ar_1 might be the ordinal of the same cardinality as Th_1). Thus, there is a one-one correspondence, ϕ', between Th_1 and Ar_1. A* is now constructed as in Th.1, using ϕ' for ϕ. Again, A and A* are isomorphic. Hence A* is a model of $\mathscr{T}$.

It follows from Th.2 that if a theory $\mathscr{T}$ has a model which provides an empirical interpretation for the theoretical terms, there is another model which assigns classes of numbers to the theoretical terms of $\mathscr{T}$. Again, the interpretation of the observation terms remains the same as do the observational relata of the mixed terms. In short, theories reconstructed within the framework given here cannot serve to characterise theoretical entities as physical, correspondence postulates notwithstanding. To see why this is so, it might be well to take a closer look at the correspondence postulates themselves.

I suspect that the main reason for the continued plausibility of the doctrine of partial interpretation of theoretical terms is a failure to recognise its incompatibility with the assumption that the theoretical and observational entities are always distinct, i.e. are disjoint classes. This can be seen from a simple example. Let 'T_1' be a theoretical term and 'O_1' an observation term (with a fixed interpretation). Now consider the following sentence:

$$(x)(T_1x \supset O_1x) \qquad (1)$$

Now the claim is typically made for sentences such as (1) that, since 'O_1' is assumed to have a fixed interpretation, (1) partially interprets 'T_1'; (1) tells us that 'T_1' refers to *some* subset of the O_1's, even though (1) in itself does not single out any particular subset (cf. [6], p. 53). Notice that coming up with any (non-empty) numerical interpretation of 'T_1' is, of course, out of the question if (1) is to be true. Since we have assumed that

---

disjoint. Then let the class assigned to 'electron' be T_i, and let q in the proof be any member of the class assigned to 'neutron'. The result is a model of the theory in which the term 'electron' now contains in its extension at least one entity (q) which was originally called a 'neutron'.

O_1 is a class of observational entities and T_1 is, by (1), a subset of O_1, T_1 must be a class of physical entities, and, in particular, observational entities. But it is here that we have the incompatibility with the assumption that the theoretical and observational entities are distinct; sentences of a form such as (1) can never serve to partially interpret theoretical terms, because, if they are to be true, then classes of theoretical entities must be subsets (or super-sets) of classes of observational entities. Actually, for reconstructed theories of the sort under consideration here, the correspondence rules would probably be similar in form to:

$$(x)(Th_x \,.\, T_1 x \,.\supset (\exists y)(Ob_y \,.\, O_1 y))^1 \qquad (1)'$$

Now that (1) has been replaced by (1)', it can be seen that (1)' allows for the pernicious juggling of interpretations as in theorems 1 and 2 above. We can now tamper with the interpretations of the theoretical terms and not disturb the interpretations of the observation terms, and this is only possible if we assume the distinctness of observational and theoretical entities. Once this assumption is made theoretical terms are then most naturally defined as those which designate classes (or properties) of theoretical entities (elements of Th) only; observation terms are those which designate classes (or properties) of observational entities (elements of Ob) only; and the mixed terms designate classes of theoretical or observational entities (elements of Th$\cup$Ob). Now it is often the case that the theoretical term—observation term distinction is not drawn so as to rule out the possibility that theoretical terms and observation terms apply to the same objects.[2] I shall call a distinction which does not rule out this possibility a 'theoretical$_1$ term'—'observation$_1$ term' distinction. In this case, the objection against correspondence rules such as (1) does not apply in general, for in my terminology, these would specify partial interpretations of *observation terms*. But the objection can be raised all over again for those of the theoretical$_1$ terms which are theoretical terms in my sense (e.g. 'electron', 'neutron', etc.). For these terms, correspondence rules such as (1) again won't do, and surely such terms would occur in any plausible reconstruction of most physical theories.

But if all this be granted, and one is sympathetic to the realist account of scientific theories, there must be something amiss somewhere. One possibility is to require that the theoretical entities be restricted to entities which are in space-time, thus ruling out numbers as possible candidates. But if we are not to abandon the notion of the implicit definition of theoretical terms altogether, then this is no significant restriction on theoretical

[1] The necessity for adopting a form such as (1)' was first pointed out to me by Grover Maxwell, and is implicit in (1).

[2] Pap ([6], p. 56) and Carnap ([1], pp. 47-48) rule this out; Hempel ([4], p. 78) does not.

entities at all. For to say that a set of entities are in space-time would merely amount to asserting that the set satisfies the *geometrical* axioms of our theory, and it follows from *Th.* 2 that a suitable class of numbers will do this for *any* set of axioms.

There is, I think, a more plausible reason which might be given as an explanation for the result given here. Theoretical entities are certainly considered as having *causal* connections with observational entities. Now the language $\mathscr{L}$ in which the theory was formulated was assumed to be extensional,[1] and perhaps our result again gives evidence to the claim that an extensional language $\mathscr{L}$ cannot 'capture' the sense of a causal connection. This contention is impossible to evaluate with any assurance, however, until a plausible and systematic account of the causal modalities appears.[2]

University of Hawaii

REFERENCES

1. Rudolf Carnap, 'The Methodological Character of Theoretical Concepts' in Feigl and Scriven, *Minnesota Studies in the Philosophy of Science*, vol. i.
2. Rudolf Carnap, *Philosophical Foundations of Physics.*
3. Carl G. Hempel, *Fundamentals of Concept Formation in Empirical Science.*
4. Carl G. Hempel, 'The Theoretician's Dilemma" in Feigl, Scriven, and Maxwell, *Minnesota Studies in the Philosophy of Science*, vol. ii.
5. Ernest Nagel, *The Structure of Science.*
6. Arthur Pap, *An Introduction to the Philosophy of Science.*

[1] $\mathscr{L}$ was also first-order. But it is easy to show that *Th.*2 (and *Th.*1) holds for all higher order languages as well.

[2] As an example of the beginning of such a program, cf. 2, pp. 208-15.