THE JOURNAL OF PHILOSOPHY

VOLUME LXVII, NO. 13, JULY 9, 1970

HOW TO DEFINE THEORETICAL TERMS

MOST philosophers of science agree that, when a newly proposed scientific theory introduces new terms, we usually cannot define the new terms using only the old terms we understood beforehand. On the contrary, I contend that there is a general method for defining the newly introduced theoretical terms.

Most philosophers of science also agree that, in order to reduce one scientific theory to another, we need to posit bridge laws: new laws, independent of the reducing theory, which serve to identify phenomena described in terms of the reduced theory with phe nomena described in terms of the reducing theory. On the contrary, I deny that the bridge laws must be posited independently. They may follow from the reducing theory, via the definitions of the theoretical terms of the reduced theory. In such cases it would be wrong to think that theoretical reduction is done voluntarily, for the sake of parsimony, when the reduced and reducing theories are such as to permit it. Sometimes reduction is not only possible but unavoidable.

F. P. Ramsey proposed that theoretical terms might be eliminated in favor of existentially quantified bound variables. Rudolf Carnap, in his most recent discussions of theoretical terms, has used Ramsey's method to split any term-introducing theory into two parts: an analytic postulate to partially interpret the theoretical terms, and a synthetic postulate in which these terms do not occur. My proposal will be in the spirit of Ramsey's and Carnap's.

My proposal could be called an elimination of theoretical terms, if you insist; for to define them is to show how to do without them. But it is better called a vindication of theoretical terms; for to define them is to show that there is no good reason to want to do without them. They are no less fully interpreted and no less well understood than the old terms we had beforehand.

I am not planning to define theoretical terms within an observation language, whatever that is. Some statements report observations, and some do not; but I do not know of any special compartment of our language that is reserved for the reporting of observations. I do not understand what it is just to be a theoretical term, not of any theory in particular, as opposed to being an observational term (or a logical or mathematical term).[1] I believe I do understand what it is to be a *T-term:* that is, a theoretical term introduced by a given theory T at a given stage in the history of science. If so, then I also understand what it is to be an *O-term:* that is, any *other* term, one of our *original* terms, an *old* term we already understood before the new theory T with its new T-terms was proposed. An O-term can have any epistemic origin and priority you please. It can belong to any semantic or syntactic category you please. Any old term can be an O-term, provided we have somehow come to understand it. And by understand I mean "understand"—not "know how to analyze."

I am also not planning to "dispense with theoretical entities." Quite the opposite. The defining of theoretical terms serves the cause of scientific realism. A term correctly defined by means of other terms that admittedly have sense and denotation can scarcely be regarded as a mere bead on a formal abacus. If it purports to name something, then if the theory that introduced it is true it does name something. I suppose a theoretical entity is something we believe in only because its existence, occurrence, etc. is posited by some theory—especially some recent, esoteric, not-yet-well-established scientific theory. Theoretical entities might better be called (as they sometimes are called) *hypothetical* entities. Theoretical terms need not name theoretical entities: consider 'H_2O'. Theoretical entities need not be named by theoretical terms: consider "living creature too small to see." Theoretical entities need not be invisible, intangible, etc.: consider the dark companions of stars. Theoretical entities are not entities of a special category, but entities we know of (at present) in a special way.

THE POSTULATE OF T

Suppose the best scientific explanation we can devise for some body of data includes a new theory T, formulated by means of a postulate in which there occur some new terms $\tau_1 \ldots \tau_n$, terms we have never used before. We shall call these newly introduced terms the *theoretical terms of* T, or just *T-terms;* and we shall call all the other

[1] My reasons are more or less those discussed by Hilary Putnam in "What Theories Are Not," *Logic, Methodology and Philosophy of Science* (Stanford: University Press, 1962), ed. by E. Nagel, P. Suppes, and A. Tarski.

terms of our scientific vocabulary *O-terms*. We shall accordingly call any sentence of our language that is free of T-terms an *O-sentence*. Let us assume that the O-terms have conventionally established standard interpretations, well known to us. The T-terms, on the other hand, are unfamiliar. Our only clue to their meaning is the postulate of T that introduced them. We are accustomed to say that it implicitly defines them; but we would prefer explicit definitions.

We may stipulate that the postulate of T is a single sentence; if it was a set of sentences, take their conjunction. If it was a finite set, we can take their conjunction within ordinary logic. If it was a decidably infinite set, we must introduce devices for infinite conjunction—to do so would be bothersome, but not problematic.

We may stipulate that our T-terms are names, not predicates or functors. No generality is lost, since names can purport to name entities of any kind: individuals, species, states, properties, substances, magnitudes, classes, relations, or what not. Instead of a T-predicate 'F __', for instance, we can use '__ has F-hood'; 'F-hood' is a T-name purporting to name a property, and '__ has __' is an O-predicate. It is automatic to reformulate all T-terms as names, under the safe assumption that our O-vocabulary provides the needed copulas:

__ has the property __
__ is in the state __ at time __
__ has __ to degree __

and the like. We will later replace T-terms by bound variables; by making the T-terms grammatically uniform, we avoid the need to introduce variables of diverse types.

We must assume that all occurrences of T-terms in the postulate of T are purely referential, open to existential generalization and to substitution by Leibniz's law. We need not assume, however, that the language of T is an extensional language. Among the O-terms there may be nonextensional operators, for instance 'it is a law that __'; nonextensional connectives, for instance '__ because __'; and so on.

We must assume, finally, that the postulate of T will be false in case any of the T-terms is denotationless. This is not a legitimate assumption for sentences in general: "There is no such substance as phlogiston" is true just because 'phlogiston' is denotationless. However, it does seem to be a legitimate assumption for the postulate of a term-introducing theory. The postulate of T will therefore imply, for each T-term τ_i, the sentence '$(\exists x)(x=\tau_i)$', which says that τ_i names something.

Such sentences sometimes count as logical truths. Many systems of logic avoid the difficulties of denotationless names by stipulating that an otherwise denotationless name is deemed artificially to name some arbitrarily chosen "null" individual. We must be able to take seriously the possibility of denotationless T-terms; it is worth the trouble to use a system of logic designed to tolerate them. Such a system has been given by Dana Scott[2]; its salient features are as follows.

(1) Improper descriptions and other denotationless terms are *really* denotationless: they name nothing in the domain of discourse. (The domain itself serves as a null individual for technical convenience; but it is not *in* the domain, and no term literally names it.)

(2) Atomic sentences containing denotationless terms are either true or false, depending on the predicate and other terms involved; we might, but need not, stipulate that they are always false.

(3) Identities containing denotationless terms on both sides are true; identities containing a denotationless term on one side only are false.

(4) Denotationless terms are interchangeable *salve veritate* in extensional contexts; necessarily denotationless names are interchangeable *salve veritate* even in intensional contexts.

THE RAMSEY AND CARNAP SENTENCES OF T

Let us write the postulate of our theory T in a way that exhibits the occurrences of T-terms therein: '$\mathsf{T}[\tau_1 \ldots \tau_n]$'.

If we replace the T-terms uniformly by variables $x_1 \ldots x_n$ respectively (distinct variables that do not occur there already), we get a formula which we may call the *realization formula* of T: '$\mathsf{T}[x_1 \ldots x_n]$'.

Any n-tuple of entities that satisfies this formula, under the fixed standard interpretations of its O-terms, may be said to *realize*, or to be a *realization* of, the theory T.

Therefore we recognize the postulate of T as the sentence that says that T is realized by the n-tuple of entities denoted, respectively, by the T-terms $\tau_1 \ldots \tau_n$. If so, then *a fortiori* T is realized. We can write another sentence, called the *Ramsey sentence* of T, which says only that T is realized: '$\exists x_1 \ldots x_n \mathsf{T}[x_1 \ldots x_n]$'.

We can write a third sentence, called the *Carnap sentence* of T, which is neutral as to whether T is realized, but says that if T is realized, then the n-tuple of entities named respectively by $\tau_1 \ldots \tau_n$ is one realization of T. The Carnap sentence is the conditional of the Ramsey sentence and the postulate:

$$\exists x_1 \ldots x_n \mathsf{T}[x_1 \ldots x_n] \supset \mathsf{T}[\tau_1 \ldots \tau_n]$$

[2] "Existence and Description in Formal Logic," in *Bertrand Russell: Philosopher of the Century* (London: Allen & Unwin, 1967), ed. by Ralph Schoenman.

Our three sentences are related as follows: (1) The postulate is logically equivalent to the conjunction of the Ramsey sentence and the Carnap sentence. (2) The Ramsey sentence and the postulate logically imply exactly the same O-sentences. (3) The Carnap sentence logically implies no O-sentences except logical truths.

Therefore, insofar as the theory T serves as a device for systematizing O-sentences, the Ramsey sentence of T will do the job as well as the postulate itself. That was Ramsey's observation.[3] The Ramsey sentence can obviously do nothing to help interpret the T-terms, since they do not occur in it. The Carnap sentence, on the other hand, does nothing to help systematize O-sentences; but it does contain the T-terms, and it does seem to do as much toward interpreting them as the postulate itself does. And the Ramsey and Carnap sentences between them do exactly what the postulate does.

Accordingly, Carnap proposes[4] to take the Ramsey sentence as the synthetic postulate of T and the Carnap sentence as the analytic postulate of T. They divide the labor of the original postulate, which both systematized O-sentences and partially interpreted the T-terms. (Here and henceforth, when I speak of Carnap's proposal it should be understood that I mean Carnap's proposal *minus* Carnap's stipulation that the O-terms belong to an observation language.)

THE INTERPRETATION OF T-TERMS

Let us see whether we want to agree that the Carnap sentence does specify the appropriate interpretations of the T-terms, insofar as appropriate interpretations can be specified at all. Put aside, for the time being, Carnap's idea that the Ramsey and Carnap sentences in partnership should be a perfect substitute for the original postulate.

It is important to separate three cases. T may have precisely one realization, or no realization, or more than one realization.

In case T is uniquely realized, the Carnap sentence clearly gives exactly the right specification. It says that the T-terms name the entities in the n-tuple that is the unique realization of T. The first T-term, τ_1, names the first component of the unique realization of T; τ_2 names the second component; and so on.

[3] "Theories," *The Foundations of Mathematics* (London: Routledge & Kegan Paul, 1931), ed. by R. B. Braithwaite.

[4] *Philosophical Foundations of Physics* (New York: Basic Books, 1966), ed. by Martin Gardner, ch. 28; "Replies and Systematic Expositions," in *The Philosophy of Rudolph Carnap* (La Salle, Ill.: Open Court, 1963), ed. by P. A. Schilpp, section 24D; "On the Use of Hilbert's ϵ-Operator in Scientific Theories," in *Essays on the Foundations of Mathematics, Dedicated to A. A. Fraenkel on his Seventieth Anniversary* (Jerusalem: Magnes Press, 1961), ed. by Y. Bar-Hillel *et al.*; "Beobachtungssprache und theoretische Sprache," *Logica: Studia Paul Bernays dedicta* (Neuchâtel: Griffon, 1959).

In case T is not realized, the Carnap sentence says nothing about the denotation of the T-terms. But this modesty seems to be uncalled for. The T-terms were introduced on the assumption that T was realized, in order to name components of a realization of T. There is no realization of T. Therefore they should not name anything. 'Phlogiston' presumably is a theoretical term of an unrealized theory; we say without hesistation that there is no such thing as phlogiston. What else could we possibly say? Should we say that phlogiston is something or other, but (unless phlogiston theory turns out to be true after all) we have no hope of finding out what?

Let us say, then, that the theoretical terms of unrealized theories do not name anything. That will do very well, at least in the case of a theory like phlogiston theory which comes nowhere near being realized. It will not do so well in the case of an unrealized theory with a (unique) *near*-realization: that is, an n-tuple that does not realize the original theory, but does realize some theory obtained from it by a slight weakening or a slight correction. We might want to say that the theoretical terms name the components of whichever n-tuple comes nearest to realizing the theory, if it comes near enough. We will ignore this complication, in part for the sake of simplicity and in part because we might hope to handle it as follows. Given a theory T, we might find a slightly weaker T', implied by but not implying T, such that an n-tuple is a realization of T' if and only if it is a near-realization of T. Then we could say that T', not T, is the real term-introducing theory; everything we have been saying about T really ought to be taken as applying to T' instead. T itself may be recovered as the conjunction of T' with further hypotheses containing the theoretical terms already introduced by T'.

There remains the case in which T is multiply realized. In this case, the Carnap sentence tells us that the T-terms name the components of some realization or other. But it does not tell us which; and there seems to be no nonarbitrary way to choose one of the realizations. So either the T-terms do not name anything, or they name the components of an arbitrarily chosen one of the realizations of T. Either of these alternatives concedes too much to the instrumentalist view of a theory as a mere formal abacus. Neither does justice to our naive impression that we understand the theoretical terms of a true theory, and without making any arbitrary choice among realizations. We should not accept Carnap's treatment in this case if we can help it. Can we?

We might say instead that the theoretical terms of multiply realized theories do not name anything. If multiple realization is a defect that theorists can reasonably hope to avoid, then we can afford to

treat multiply realized theories as failures: call them false, and call their theoretical terms denotationless. But if multiple realization is inevitable, we cannot afford to disdain multiply realized theories. We can have denotations arbitrarily chosen, or no denotations at all.

A uniquely realized theory is, other things being equal, certainly more satisfactory than a multiply realized theory. We should insist on unique realization as a standard of correctness unless it is a standard too high to be met. Is there any reason to think that we must settle for multiply realized theories? I know of nothing in the way scientists propose theories which suggests that they do not hope for unique realization. And I know of no good reason why they should not hope for unique realization. Therefore I contend that we ought to say that the theoretical terms of a multiply realized theories are denotationless.

Many philosophers do seem to think that unique realization is an extravagant hope, unlikely in scientific practice or even impossible in principle. Partly, this is professional skepticism; partly it is skepticism derived from confusions that I shall try to forestall.

In the first place, I am not claiming that scientific theories are formulated in such a way that they could not possibly be multiply realized. I am claiming only that it is reasonable to hope that a good theory will not in fact be multiply realized.

In the second place, I am not claiming that there is only one way in which a given theory *could* be realized; just that we can reasonably hope that there is only one way in which it *is* realized.

Finally, I should say again that we are talking only about realizations that make T true under a fixed interpretation of all of its O-vocabulary. And this O-vocabulary may be as miscellaneous as you please; in practice it is likely to be very miscellaneous indeed. An O-term is *any* term, of any character, which we already understood before the new theory T came along. It does not have to belong to an observation language. If anyone hopes to adapt my proposals to the task of interpreting theoretical terms using only an observation language—if there is any such thing—I would not be at all surprised if he ran into trouble with multiple realizations. But his project and his troubles are not mine.

John Winnie has announced a proof that scientific theories cannot be uniquely realized.[5] Though his proof is sound, it goes against nothing I want to say. Most of Winnie's multiple realizations of a given theory—all but one, perhaps—are not what I call realizations of the theory. I am concerned only with realizations

[5] "The Implicit Definition of Theoretical Terms," *British Journal for the Philosophy of Science*, XVIII (1967): 223–229.

under a fixed interpretation of the O-vocabulary; whereas Winnie permits variation in the interpretation of certain O-terms from one realization to another, provided that the variation is confined to theoretical entities. For instance, he would permit variation in the extension of the O-predicate '__ is bigger than __' so long as the extension among observational entities remained fixed. Winnie's proof does not show that a theory is multiply realized in my sense unless the postulate of the theory is free of "mixed" O-terms: O-predicates whose extension includes theoretical entities, and the like. I would claim that mixed O-terms are omnipresent, and that there is no reason not to grant them a fixed interpretation even as applied to theoretical entities.

Perhaps another reason to think that theories cannot be uniquely realized comes from the idea that theoretical terms are somehow partially interpreted. It seems that the stronger a theory is, the better its theoretical terms are interpreted. If the postulate of the theory is a tautology, for instance, the theoretical terms are not interpreted at all. It is tempting to explain this by saying that the stronger theory has fewer realizations. Since no consistent theory interprets its terms so well that it could not have done better if it had been still stronger, it seems that unique realization—perfect interpretation—is a limit we can never reach. But this is a mistake. The stronger theory *may* have fewer actual realizations or it may not; but it *must* have less *risk* of multiple realization, and that is enough to explain why strength seems to make for better interpretation. On the other hand, the stronger theory must also have more risk of nonrealization.

Let us conclude, therefore, that the T-terms ought to name the components of the unique realization of T if there is one, and ought not to name anything otherwise. We can record our conclusion by laying down three meaning postulates. The first

$$\exists y_1 \ldots y_n \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv . y_1 = x_1 \& \ldots \& y_n = x_n) \supset \top[\tau_1 \ldots \tau_n]$$

says that if T is uniquely realized, then it is realized by the entities named, respectively, by $\tau_1 \ldots \tau_n$. It is logically implied by the Carnap sentence of T. The second

$$\sim\exists x_1 \ldots x_n \top[x_1 \ldots x_n] \supset . \sim\exists x(x = \tau_1) \& \ldots \& \sim\exists x(x = \tau_n)$$

says that, if T is not realized, then $\tau_1 \ldots \tau_n$ do not name anything. It is logically independent of the Carnap sentence. The third

$$\exists x_1 \ldots x_n \top[x_1 \ldots x_n] \& \sim\exists y_1 \ldots y_n \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv . y_1 = x_1 \& \ldots \& y_n = x_n) . \supset . \sim\exists x(x = \tau_1) \& \ldots \& \sim\exists x(x = \tau_n)$$

says that, if T is multiply realized, then, again, $\tau_1 \ldots \tau_n$ do not

name anything. It disagrees with the Carnap sentence, inasmuch as the third postulate and the Carnap sentence together imply that T has at most one realization, but that conclusion ought not to follow from meaning postulates.

Now we have specified the denotations of the T-terms. What about their senses? We have specified their senses already. For we have specified their denotations in any possible world, not just here in our actual world. In any possible world, they are to name the components of whatever uniquely realizes T in that world, and they are to name nothing in that world unless T is uniquely realized there. We know what it is for an n-tuple of entities to realize T in an arbitrary possible world w: namely, the n-tuple satisfies the realization formula of T in the model determined jointly by the state of affairs in the possible world w and by the fixed interpretation of the O-vocabulary.

Here I rely on, but do not argue for, a doctrine of senses due originally to Carnap.[6] I am supposing that the sense of a name—at least, of a name that cannot be decomposed into constituents—is given in full by specifying what (if anything) it names in each possible world. If we like, we can say that a sense *is* a function from (some or all) possible worlds to named entities. The most important objection to this doctrine is that possible worlds and their unactualized inhabitants are occult; I have argued elsewhere[7] that they are no more occult than the infinite sets we have learned to live with, and just as useful in systematic philosophy.

Putnam, arguing against Carnap's notation of partial interpretation, has objected that "theories with false observational consequences have *no* interpretation (since they have no model that is 'standard' with respect to the observation terms). This certainly flies in the face of our usual notion of an interpretation, according to which such a theory is *wrong*, not *senseless*."[8] The objection is mistaken. The theoretical terms of such a theory are not senseless, just *denotationless*. Their sense is given by their denotation in those possible worlds in which the theory *is* uniquely realized and, therefore, does *not* have false consequences. They have just as much sense as the denotationless term 'Santa Claus'.

A *logically determinate* name is one which names the same thing in every possible world. Its sense is a constant function. Numerals, for instance, seem to be logically determinate names of numbers.

[6] *Meaning and Necessity*, 2nd ed. (Chicago: University Press, 1956), pp. 181–182.

[7] *Convention* (Cambridge, Mass.: Harvard, 1969), p. 208.

[8] "What Theories Are Not," p. 247.

But 'the number of solar planets' is a logically indeterminate name of a number. Here in our actual world it happens to name the number 9. Elsewhere it may name other numbers, or nothing at all. *Anything* can have a logically indeterminate name—even a property. (This should have been obvious, but wasn't.) Take the name 'the physical property detected by means of the instrument with catalog number 12345 in so-and-so catalog' (filling in the name of a catalog). This happens, perhaps, to name the property of fluorescing in the ultraviolet. But if the catalog numbers had been different, it would have named some other property, or none.

We should notice that T-terms purporting to name properties will normally turn out to be logically indeterminate names of properties. Likewise for other T-terms; but let us stick to the special case. Suppose τ_1 is a T-term purporting to name a property. More precisely: T is formulated in such a way that the postulate of T cannot be true unless τ_1 names a property. It follows that if any n-tuple of entities uniquely realizes T in any possible world, the first component of that n-tuple is a property. But unless T is a very special theory, different n-tuples with different first components will uniquely realize T in different possible worlds. The sense of τ_1 will not be a constant function.

The logical indeterminacy of τ_1 makes for subtle equivocation in any context in which possible worlds other than our actual world are under discussion, either explicitly or implicitly. For instance, there will be trouble whenever τ_1 occurs in the scope of a modal operator, in the scope of a nomological operator, in the scope of an epistemic operator, or in a subjunctive conditional. We have to keep track of when τ_1 names the first component of the unique realization of T (if any) in our actual world, and when τ_1 instead names the first component of the unique realization of T (if any) in some other possible world under discussion.

For instance, someone might plausibly object that, on my account, it is impossible for T to have a unique realization but for the first component thereof not to be the property named by τ_1. But that does seem to be possible. Just consider the property named by τ_1. Fix your attention on that property, whatever it is. Now surely T could have a unique realization in which *that* property was not the first component! So what is impossible according to me is really possible. I reply that the objection commits a fallacy of equivocation. What I assert is this: for no possible world w is it the case that T has a unique realization in w but that the first component thereof is not the property named by τ_1 in the world w. What the objector properly denies is this: for no possible world w is it the case that

T has a unique realization in w but that the first component thereof is not the property named by τ_1 in our actual world. Unfortunately, what I assert and what he properly denies can both be expressed by this ambiguous sentence: "It is impossible for T to have a unique realization but for the first component thereof not to be the property named by τ_1."

A similar difficulty arises over the name 'the property of having τ_1'. This name purports to name a property. We would suppose offhand that it names the same property as τ_1 itself. Indeed, we would suppose it has the same sense as τ_1. But at least on one reading it and τ_1 do not name the same property, neither in our actual world nor in any other possible world.

I take it that a property is identified when, and only when, we have specified exactly which things have it in every possible world. And I take it that a name of the form 'the property of doing so-and-so' names the property that belongs, in any world w, to exactly those things which, in the world w, do so-and-so. For instance, 'the property of having τ_1' names the property that belongs, in any world w, to exactly those things which, in the world w, have the property named by τ_1.

Now we can see the problem. Do we mean: (1) the property that belongs, in any world w, to exactly those things which, in the world w, have the property named by τ_1 *in our actual world?* That, of course, is just the same property which is named by τ_1 in our actual world. On this first reading, 'the property of having τ_1' and τ_1 do both name the same property.

Or do we rather mean: (2) the property that belongs, in any world w, to exactly those things which, in the world w, have the property named by τ_1 *in the world w?* On this second—and, I believe, better—reading, 'the property of having τ_1' is a logically determinate name of a certain property, which we may call the *diagonalized sense* of τ_1. The sense of τ_1 may be represented by a function $\|\tau_1\|$ which assigns to any world w a property $\|\tau_1\|_w$. A property in turn may be represented by a function P which assigns to any world w the set P_w of things which, in the world w, have the property. Then the diagonalized sense of τ_1 is the property whose representing function assigns to any world w the set of things $(\|\tau_1\|_w)_w$. It is not named by τ_1 in any world, unless T is a very peculiar theory. Neither is it the sense of τ_1; that is not a property at all, but rather a function from worlds to properties.

THE DEFINITIONS OF T-TERMS

Given our conclusion that the T-terms $\tau_1 \ldots \tau_n$ should denote the components of the unique realization of T if there is one, and should

not denote anything otherwise, it is natural to define the T-terms by means of definite descriptions as follows.

$$\tau_1 = \iota y_1 \exists y_2 \ldots y_n \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv . y_1 = x_1 \& \ldots \& y_n = x_n)$$
$$\ldots$$
$$\tau_n = \iota y_n \exists y_1 \ldots y_{n-1} \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv . y_1 = x_1 \& \ldots \& y_n = x_n)$$

These are to be our *definition sentences* for the theory T. The first, for instance, says that τ_1 names that entity which, followed by some $n-1$ entities, comprises an n-tuple identical with all and only n-tuples that realize T. That is to say, τ_1 names the first component of the unique realization of T.

These definitions work properly, under the treatment of denotationless terms we have chosen. They are valid: true in any model in which the T-terms are interpreted as specified in the previous section, whether or not T is uniquely realized therein. They are jointly equivalent to the set of meaning postulates we put forth to replace the Carnap sentence of T. They do not imply any O-sentences except logical truths.

We can see now why it was worth the trouble to adopt Dana Scott's treatment of denotationless names rather than its more familiar alternatives.[9] Under Russell's theory of descriptions, the definition sentences are disguised existential quantifications, false unless T is uniquely realized. Under the truth-value-gap theory of Frege and Strawson, atomic contexts of denotationless names or descriptions have no truth value; so the definition sentences are neither true nor false unless T is uniquely realized. Either way, the definition sentences would not be valid identities, true whether or not T is uniquely realized. We would be forced to regard them as metalinguistic assertions of synonymy, or something of the sort. Under the chosen-individual theory of Frege and Carnap, improper descriptions name some arbitrarily chosen individual. So do the theoretical terms of unrealized or multiply realized theories, if the definition sentences are to be valid. But suppose T is multiply realized, and suppose the n-tuple consisting of the chosen individual taken n times over happens to be one of the realizations of T. In this case, by accident, the postulate of T turns out to be true, contrary to our decision that multiply realized theories should be false.

THE EXPANDED POSTULATE OF T

Given our definitions, we can eliminate T-terms in favor of the definite descriptions whereby we have defined them. Replacing each

[9] For a survey of the alternatives, see David Kaplan, "What Is Russell's Theory of Descriptions?" *Physics, Logic, and History* (New York: Plenum, 1969), ed. by Wolfgang Yourgrau.

T-term by its definiens throughout the postulate of T, we obtain an O-sentence which we may call the (*definitionally*) *expanded postulate* of T. It says T is realized by the n-tuple consisting of the first, second, ..., nth components of the unique realization of T. The expanded postulate is, of course, definitionally equivalent to the postulate. That is, the postulate together with the definition sentences is logically equivalent to the expanded postulate together with the definition sentences.

The expanded postulate says that T is uniquely realized. It is logically equivalent to a shorter O-sentence that says so in a more straightforward way; we may call this the *unique-realization sentence* of T:

$$\exists y_1 \ldots y_n \forall x_1 \ldots x_n (\top [x_1 \ldots x_n] \equiv . y_1 = x_1 \& \ldots \& y_n = x_n)$$

Thus the postulate is definitionally equivalent to the unique-realization sentence. That is what we should expect, given our decision—*contra* Carnap—to interpret the T-terms in such a way that the postulate is true if and only if T is uniquely realized.

Still we may have misgivings. The expanded postulate is an O-sentence stronger than the Ramsey sentence of T, which said merely that T had *at least* one realization. Yet if the definition sentences are part of T, the expanded postulate is an O-theorem of T. So the definitions are giving us O-theorems that could not have been derived without them. That means that the definitions themselves, unlike the Carnap sentence, are not logically implied by the postulate.

Therefore, if I want to contend that the definition sentences of T are correct definitions, I must give up the idea that the theorems of T are all and only the logical consequences of the postulate of T. I am quite willing to give up that idea. I contend that the theorist who proposed T by asserting the postulate of T explicitly, labeling it as the postulate of a term-introducing theory, has also implicitly asserted the definition sentences of T. I contend that his audience *will* take him to have implicitly asserted the definition sentences of T. That is an empirical hypothesis about the conventional semantics of our language. To test it, we should find out what would happen if the audience came to think that T was multiply realized.

If they would thereupon call T a false theory, that confirms my account of the interpretation and definitions of the T-terms. If they would call T a true theory, that refutes my account and confirms Carnap's. What if they would be unable to decide which to say? Then my account and Carnap's are rival proposals for rational reconstruction within the range of free choice left open by our conventions. My proposal has the advantage that it permits theoretical

terms to be fully interpreted and explicitly defined. Carnap's proposal has the advantage that it lets us live with multiply realized theories. But I see no reason why we need to be able to live with multiply realized theories.

The expanded postulate (and its logical equivalent, the unique-realization sentence) are not by any means the only interesting definitional equivalents of the postulate of T. For any T-term τ_i, consider the sentence that says τ_i names something: '$(\exists x)(x=\tau_i)$'. This sentence is definitionally equivalent to the postulate of T and to the expanded postulate of T (but logically equivalent to neither). That is what we should expect: T is uniquely realized if and only if there is something which is the ith component of the unique realization of T.

Therefore the postulate and expanded postulate of T are definitionally implied by any sentence which contains T-terms and which comes out false in case any of its T-terms are denotationless. There are many such sentences. Let τ_1, τ_2, and τ_3 be T-terms purporting to name, respectively, a property, a relation, and a function; and let ζ and η be O-terms purporting to name individuals. Let ν be an O-term naming a real number. Now consider such sentences as: 'ζ has the property τ_1', 'ζ bears the relation τ_2 to η', or 'the value of the function τ_3 for the argument ζ is ν'. Clearly each of these sentences should be false if its T-term is denotationless. Therefore each of them definitionally implies the postulate and expanded postulate of T.

We would probably say, at first sight, that sentences such as these purport to state particular facts. Yet the theoretical postulates they imply—or some conjuncts thereof—will usually purport to state laws of nature, perhaps even laws of unrestricted universality. Therefore it is possible that a particular fact might be explained by a covering-law explanatory argument in which the explanans contain no explicit statement of any law! The required covering laws would be available by implication from certain particular-fact premises in the explanans. It seems that, if my account of theoretical terms is correct, the covering-law analysis of scientific explanation is in need of a slight reformulation. Take one of Hempel's definitions of deductive-nomological explanation:

> ...deductive arguments whose conclusion is the explandum sentence...; and whose premiss-set, the explanans, consists of general laws..., and of other statements..., which make assertions about particular facts.[10]

[10] *Philosophy of Natural Science* (Englewood Cliffs, N.J.: Prentice-Hall 1966), p. 51.

I contend we ought rather to say:

> ...deductive arguments whose conclusion is the explanandum sentence; and whose premiss-set, the explanans, consists of statements which make assertions about particular facts and *perhaps* also of statements which assert general laws; and whose premiss-set implies statements which assert general laws.

I am confident that my objection goes against the letter, not the spirit, of Hempel's covering-law analysis.

DERIVED BRIDGE LAWS FOR T

We can safely assume that there is a period, after the new theory T has been proposed, during which the T-terms retain the interpretations they received at the time of their introduction. At least for a while, our definition sentences for T remain valid. Suppose that, during this period, T is reduced by means of some other accepted scientific theory T^*. Let this be a reduction in which T survives intact; not, what is more common, a simultaneous partial reduction and partial falsification of T by T^*. We shall consider how this reduction might take place.

The reducing theory T^* need not be what we would naturally call a single theory; it may be a combination of several theories, perhaps belonging to different sciences. Parts of T^* may be miscellaneous unsystematized hypotheses which we accept, and which are not properly called theories at all. Different parts of T^* may have been proposed or accepted at different times, either before or after T itself was proposed.

The most interesting case, however, is that in which T^* is well systematized, and at least part of T^* is newer than T. It is in that case that the reduction of T by means of T^* is likely to be an important advance toward the systematization of all empirical knowledge.

T^*, or parts of T^*, may introduce theoretical terms; if so, let us assume that these T^*-terms have been introduced by means of the same O-vocabulary which was used to introduce the theoretical terms of T. This is possible regardless of the order in which T and T^* were proposed. Any term that is either an O-term or a T^*-term may be called an O^*-term; so at the time T is reduced, the relevant part of our scientific vocabulary is divided into the T-vocabulary and the O^*-vocabulary.

Suppose the following O^*-sentence is a theorem of T^*; we may call it a *reduction premise* for T: '$\top[\rho_1 \ldots \rho_n]$'. The terms $\rho_1 \ldots \rho_n$ are to be names belonging to the O^*-vocabulary. They may be elementary expressions and belong to the O^*-vocabulary in their own right, or they may be compound expressions—for instance, definite descrip-

tions—whose ultimate constituents belong to the O^*-vocabulary. The reduction premise says that T is realized by an n-tuple of entities named, respectively, by the O^*-terms $\rho_1 \ldots \rho_n$. Notice that it cannot be true if any of those O^*-names are denotationless.

A reduction premise for T does not imply the postulate of T either logically or definitionally; nor conversely. But the postulate does follow logically from the reduction premise together with a set of *bridge laws* for T, as follows:

$$\rho_1 = \tau_1$$
$$\ldots$$
$$\rho_n = \tau_n$$

The bridge laws serve to identify phenomena described in terms of the reduced theory T with phenomena described in terms of the reducing theory T^* (including the O-vocabulary); and via these identifications, T can be derived from T^*.

But where do we get the bridge laws? The usual view[11] is that they are separate empirical hypotheses, independent of the reducing theory T^*. When T^* yields a reduction premise for T, we have the opportunity to choose between two rival bodies of theory. We may choose T^* augmented with bridge laws; if so, we can derive T, so we do not have to posit it. Or we may choose T^* without bridge laws, in which case we will have to posit T separately. Given this choice, we take whichever body of theory is better—more systematized, parsimonious, simple, credible, or what have you. We must decide whether the gain in systematization from reducing T is worth the loss in systematication from adding bridge laws. If it is, we choose to accept the bridge laws and perform the reduction.

The usual view assumes that it is impossible to derive the bridge laws from the unaugmented reducing theory T^*, since the bridge laws contain essential occurrences of the T-terms and these do not occur in T^*. But consider the *definitionally expanded bridge laws:* the sentences obtained from the bridge laws when we replace the T-terms by definite descriptions, according to our definition sentences for T.

$$\rho_1 = \iota y_1 \exists y_2 \ldots y_n \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n)$$
$$\ldots$$
$$\rho_n = \iota y_n \exists y_1 \ldots y_{n-1} \forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n)$$

The definitionally expanded bridge laws are O^*-sentences. No incompatibility of vocabulary prevents them from being theorems of

[11] Reviewed in the first two sections of John Kemeny and Paul Oppenheim, "On Reduction," *Philosophical Studies*, VII (1956): 6–13. Since Kemeny and Oppenheim take the theoretical terms as predicates, their bridge laws are universally closed biconditionals rather than identities.

T^*. Yet they are definitional equivalents of the bridge laws; so if they are theorems of T^*, then T^* definitionally implies the bridge laws. There is no need for any empirical hypothesis other than theorems of T^*.

If T^* yields as theorems a reduction premise for T, and also a suitable set of definitionally expanded bridge laws for T, then T^*—without the aid of any other empirical hypothesis—reduces T. For T^* definitionally implies the postulate of T, as well as a set of bridge laws. Once T^* is accepted, there is no choice whether or not to reduce T. The reduction of T does not need to be justified by considerations of parsimony (or whatever) over and above the considerations of parsimony that led us to accept T^* in the first place.

It is useful to observe that the set of definitionally expanded bridge laws is logically implied by the following O^*-sentence, which we may call an *auxiliary reduction premise* for T:

$$\forall x_1 \ldots x_n (\top[x_1 \ldots x_n] \equiv . \rho_1 = x_1 \& \ldots \& \rho_n = x_n)$$

It says that, unless one of the terms $\rho_1 \ldots \rho_n$ is denotationless, T is uniquely realized by an n-tuple of entities named, respectively, by $\rho_1 \ldots \rho_n$. Hence the reduction premise and the auxiliary reduction premise together say just this: T is uniquely realized by an n-tuple of entities named, respectively, by $\rho_1 \ldots \rho_n$. That is what T^* must imply in order to reduce T by means of derived bridge laws. In that case, indeed, we can by-pass the bridge laws. T^* definitionally implies T by an alternate route. Since T^* guarantees that T is uniquely realized, T^* logically implies the unique realization sentence of T, which is logically equivalent to the expanded postulate of T and definitionally equivalent to the postulate of T.

Let us briefly examine two examples of reduction by means of derived bridge laws.

(1) Let T be a theory explaining the operation of a machine by means of transition laws of the form: when the machine is in state τ_i, so-and-so input causes it to go into state τ_j and produce such-and-such output. The T-terms are $\tau_1 \ldots \tau_n$, purporting to name states of the machine. Later we obtain T^*: an account, purporting to be complete, of the internal structure of the machine and the principles on which it works. Then, unless T^* falsifies T, we should be able to form state-names $\rho_1 \ldots \rho_n$ in the mechanical vocabulary of T^*, such that T^* implies that the states thus named are related to one another, and to the appropriate inputs and outputs, by the given transition laws. Thus T^* would yield a reduction premise for T. But also, since T^* purports to be a complete account of the working of the machine, we can reasonably expect T^* to leave no room for a second realization of T. T^* could imply that every

n-tuple of states describable in its mechanical vocabulary, except for the n-tuple named by $\rho_1 \ldots \rho_n$, disobeys the given transition laws; and according to the completeness claimed by T^*, no states not so describable need be considered in explaining the input-output relations of the machine. (For instance, what we think we know about ordinary electromechanical machines seems to imply that the only states causally responsible for outputs are states describable in terms of positions and momenta of moving parts, currents, and voltages.) If so, T^* yields an auxiliary reduction premise for T; hence T^* together with the definitions of the state-names $\tau_1 \ldots \tau_n$ suffices to imply the bridge laws: '$\tau_i = \rho_i$'.

(2) Let T be a theory explaining the regulation of certain biological processes by positing hormones $\tau_1 \ldots \tau_n$: chemical substances of unspecified composition, secreted by specified cells under specified conditions and regulating the rates of specified chemical reactions in a specified way. The T-terms $\tau_1 \ldots \tau_n$, in this case, purport to name substances. Let T^* comprise our body of biochemical knowledge at some later time; T^* might imply that certain substances named by chemical formulas $\rho_1 \ldots \rho_n$ realized T, and that they alone did so. To exclude multiple realization of T, T^* would have to contain the information that, e.g., a certain gland secretes *nothing but* the substance with formula ρ_1; but we often do have such knowledge.

Not only is it possible for a theory to be reduced by means of derived bridge laws; we can even regard all possible reduction of theories as working in this way. Instead of saying that in some cases the bridge laws are posited independently of the reducing theory, we may rather say that in some cases the reducing theory must be strengthened ad hoc before it yields the bridge laws we want.

Suppose we want to reduce our theory T; and we accept a theory T^* which yields as a theorem some reduction premise for T, but does not yield the corresponding set of definitionally expanded bridge laws. Let T^{**} be the theory obtained from T^* by providing these definitionally expanded bridge laws. T^* does not reduce T by means of derived bridge laws, but only by means of bridge laws posited independently of T^*. T^{**}, a theory obtained by suitable strengthening of T^*, does reduce T by means of derived bridge laws. How shall we describe what happens when we add bridge laws to T^* in order to reduce T? If we say that T^* is the reducing theory, we must say that T^* reduces T by means of independently posited bridge laws. But we could just as well say that T^{**} was the reducing theory, and that we strengthened T^* to T^{**} in order that T^{**} might reduce T by means of derived bridge laws. Of course, it does not

matter very much which we say. Either way, the bridge laws have been posited ad hoc in return for the reduction of T; for T^{**} is definitionally equivalent to T^{*} plus the bridge laws.

Still, I think it best to say that the strengthening of the reducing theory is a precondition, not a part, of the reduction. For that will remind us that strengthening is not necessarily needed. A body of theory already accepted before we thought of using it to reduce T may already be strong enough, without any more strengthening ad hoc, to reduce T by means of derived bridge laws. I do not know whether most cases of theoretical reduction are of this sort; but we should at least leave the possibility open, as the usual account of reduction does not.

LATER REVISIONS OF T

So far, we have discussed the interpretation of T-terms only at the time of their introduction, the time when their parent theory T is first proposed. It remains to ask what happens later when T is amended and extended. This matters especially in connection with reduction, since theories do not often survive reduction intact. More often the original theory is falsified while a corrected version is reduced. If T is thus partially reduced and partially falsified, or revised for any other reason, do the T-terms retain their meanings?

We might say that the T-terms should always be defined using the currently accepted version of T. As T is corrected, modified, extended, or perhaps even when we accept miscellaneous hypotheses that contain T-terms but do not belong integrally to any version of T, the T-terms gradually change their meaning. In particular, they change their meanings in the revisions preparatory to partial reduction of T.[12] But these are very peculiar changes of meaning—so peculiar that this position seems to change the meaning of 'change the meaning of'. They occur continually, unnoticed, without impeding communication. We might try saying that a *small* enough change of meaning is not really a change of meaning, but that would imply that enough nonchanges in meaning could add up to a change in meaning. The position has other problems. How are we to define the theoretical terms of defunct or never-accepted theories? Should we use the best-known version, or what? What if different scientists accept slightly different versions of T, disagreeing, say, on the exact value of some physical constant? We ought not to say they give the T-terms different meanings. What do the T-terms mean if we have suspended judgment between two slightly different versions of T?

[12] As suggested by Paul Feyerabend, "Explanation, Reduction, and Empiricism," *Minnesota Studies in the Philosophy of Science* III (Minneapolis: University of Minnesota Press, 1962), ed. by H. Feigl and G. Maxwell.

We might therefore prefer to say that the T-terms keep the meanings they received at their first introduction. They should still be defined using the original version of T even after it has been superseded by revised versions.

This position will work only if we permit the T-terms to name components of the nearest near-realization of T, even if it is not a realization of T itself. For after T has been corrected, no matter how slightly, we will believe that the original version of T is unrealized. We will want the T-terms to name components of the unique realization (if any) of the corrected version of T. They can do so without change of meaning if a realization of the corrected version is also a near-realization of the original version.

According to this position, we may be unable to discover the meanings of theoretical terms at a given time just by looking into the minds of the most competent language-users at that time. We will need to look at the past episodes of theory-proposing in which those terms were first introduced into their language. The working physicist is the expert on electrons; but the historian of physics knows more than he about the meaning of 'electron', and hence about which things could truly have been called electrons if the facts had been different. If we were ignorant of history, we could all be ignorant or mistaken about the meanings of words in common use among us. This situation is surprising, but it has precedent: a parallel doctrine about proper names has recently been defended.[13] To know what 'Moses' means among us it is not enough to look into our minds; you must look at the man who stands at the beginning of the causal chain leading to our use of the word 'Moses'.

I do not wish to decide between these alternatives. Either seems defensible at some cost. I hope the truth lies in between, but I do not know what an intermediate position would look like.

DAVID LEWIS

Princeton University

EXPLANATION AND REDUCTION *

ON what has come to be a standard account of the matter, to *explain* some phenomenon is to derive a sentence describing it from other sentences at least one of which states a general law. Thus a deductive argument whose premises include the law for reflection in a plane mirror and a description of certain

[13] See David Kaplan, "Quantifying In," *Synthese*, XIX (1968): 178–214.

* This is a revised version of a paper read at the Pacific Division meeting of the American Philosophical Association, March, 1969. John Earman, Karel Lambert, and John Vickers have supplied helpful comments.