Theory-Dependent Terms

Author(s): David Papineau

Source: *Philosophy of Science*, Mar., 1996, Vol. 63, No. 1 (Mar., 1996), pp. 1-20

Published by: The University of Chicago Press on behalf of the Philosophy of Science Association

Stable URL: https://www.jstor.org/stable/188222

Philosophy of Science

March, 1996

THEORY-DEPENDENT TERMS*

DAVID PAPINEAU†‡

Department of Philosophy
King's College London

The main puzzle about theoretical definitions is that nothing seems to decide which assumptions contribute to such definitions and which do not. I argue that theoretical definitions are indeed imprecise, but that this does not normally matter, since the definitional imprecision does not normally produce indeterminacy of referential value. Sometimes, however, the definitional imprecision is less benign, and does generate referential indeterminacy. In these special cases, but not otherwise, it is necessary to refine the term's definition.

1. Introduction. It is a familiar idea that the meanings of some words derive from their place in a theory. So, for example, the meanings of words for scientific unobservables have been widely argued to gain their significance from the way that scientific theories link them with other such words and with words for observables. Similarly, the meanings of various everyday words, such as the words used in everyday psychology, are often held to derive their meanings from common sense theories like "folk psychology."

However, the idea of such definitions[1] is not unproblematic. Exactly which assumptions are supposed to contribute to theoretical definitions?

*Received December 1994; revised January 1995.

†For help with this paper, I would like to thank Tad Brennan, Tim Crane, Michael Devitt, Keith Hossack, Sarah Patterson, Stathis Psillos, Murali Ramachandran, Mark Sainsbury, Barry Smith, Steven Stich, Scott Sturgeon, and Bernhard Weiss.

‡Send reprint requests to the author, Department of Philosophy, King's College, London, England WC2R 2LS.

[1]This terminology of "definitions" should not be taken to imply that some person once successfully stipulated a definition for the "defined" term. I assume only that "theoretically defined" terms have just the meanings that they would have if they were governed by such stipulations.

Philosophy of Science, 63 (March 1996) pp. 1–20. 0031-8248/96/6301-0001$2.00

Quinean considerations suggest that there is no way of drawing a line between analytic assumptions that play a defining role and synthetic assumptions that do not. Certainly there is no obvious feature of scientific or everyday thinking which might serve to underpin such a distinction. But this then threatens the implication that the meaning of theory-dependent terms is imprecise and that claims made using them are therefore not well-defined.

In this paper I shall argue that the meanings of theory-dependent terms are indeed imprecise, but that this does not normally matter. This is because the imprecision in definition does not normally lead to an indeterminacy in referential value.

I say that the imprecision of theoretical definitions does not *normally* lead to an indeterminacy in referential value, and so does not *normally* matter. The qualification is non-trivial, in that there are *some* theory-dependent terms which do have indeterminate referential values. When this less benign species of imprecision is detected, the appropriate remedy is to remove it by tightening up the relevant term's definition.

2. Related Issues. The imprecision of theoretical definitions bears on a number of recent philosophical debates. Most obviously, it is relevant to recent discussions of "semantic holism." One familiar argument for semantic holism starts with the assumption

(1) the meanings of some terms are fixed by theoretical definitions.

It then adds in the Quine-inspired premise

(2) we cannot divide the assumptions in a theory into those with definitional status and those without.

And from these two premises it concludes

(3) *all* the assumptions containing a theoretically defined term contribute equally to its meaning.

Since most philosophers want to resist this conclusion, they deny one or another of the premises. Some, like Fodor, object to the first, denying any possibility of terms whose meanings are determined by their theoretical role (Fodor 1987, 73–94). Others, like Devitt, resist the second assumption and seek some sharp distinction between meaning-constituting assumptions and others (Devitt, forthcoming, Ch. 3). I find neither of these options plausible. The view that theory-dependent concepts exhibit a harmless species of imprecise definition will cast a new light on this issue.

In the last three or four years, a more specific debate in the philosophy of psychology has also focused interest on the topic of this paper. Connectionist models of mind suggest that our cognitive structure lacks some of the features commonly ascribed to it. For example, connectionist models suggest that no cognitive states have the kind of internal causal structure that beliefs are widely assumed to possess. Now, does this mean that

connectionism implies that there are no *beliefs*? If "belief" is defined by a set of common sense psychological assumptions, as many contemporary philosophers of mind suppose, then the answer to this question hinges on whether the assumption challenged by connectionism—that beliefs have such-and-such internal causal structure—is a member of this defining set. That is, connectionism implies that there are no beliefs if and only if "beliefs have such-and-such internal causal structure" is part of the definition of "belief." But what determines whether this assumption is included in the definition of "belief" or not?

Stich (together with Ramsey and Garon) famously argued in "Connectionism, Eliminativism and the Future of Folk Psychology" (1990) that if there are indeed no cognitive states with the relevant internal causal structure, then there are no beliefs. But more recently, doubts about the meaning of "belief" have made Stich more cautious. In a later paper (1991), he argues that there is no fact of the matter about whether connectionism implies there are no beliefs. I shall end up agreeing with Stich that the answer to this question is indeterminate. But my position is somewhat different from his. He argues, first, that our linguistic intuitions will be indecisive in deciding what "belief" refers to and, second, that there is nothing especially important about the reference relation picked out by our linguistic intuitions anyway. The overall argument of this paper can be viewed as one way of elaborating the first of these thoughts, but it lends no support to the second.

So far I have focused on the possible imprecision of theoretical definitions. But there is also another philosophical worry raised by theoretical definitions. If we define a word in terms of some theory, doesn't this make the theory analytic? And doesn't this reduce decisions on whether or not to accept the theory to choices of convention? For example, if the meaning of F is partly fixed by the assumption that "All Fs are Gs," then doesn't this make "All Fs are Gs" analytic, and any decision to alter it a mere linguistic ruling?[2]

This issue was widely discussed a couple of decades ago, under the heading of the "problem of meaning variance" (cf. Shapere 1966, Scheffler 1967). In the 1960s and 1970s, Quine's arguments against the analytic-synthetic distinction, in combination with Kuhn's and Feyerabend's emphasis on the importance of theoretical presuppositions (Quine 1951, Kuhn 1962, Feyerabend 1962), persuaded many philosophers that scientific change is inseparable from meaning change. This in turn generated doubts about the rationality of scientific theory-choice.

Despite the widespread attention devoted to "the problem of meaning

[2]It will be convenient to use "F", "G", etc., when standing alone as dummy names of words, that is, as if they are enclosed by corner quotes. When I want a dummy name for a theoretical property or other entity, I shall use italics, as in "*F*", "*G*", etc.

variance" in those decades, no clear solution was agreed upon. The issues were not so much resolved as forgotten. This was due to the emergence of causal theories of reference in the 1970s. Though these theories were primarily designed as an account of proper names of spatio-temporal particulars, they were also applied to natural kind terms, terms for biological species, and terms for unobservable scientific properties. As in the case of proper names, the referents of these other terms were argued to be fixed, not by speakers' beliefs about the referent, but by some original occasion where (a sample or manifestation of) the referent was dubbed with the term. Later uses of the term then also referred to whichever entity had been present at the dubbing.

As a number of writers quickly observed, this model of meaning for scientific terms removes "the problem of meaning variance." Since the causal theory of reference makes meanings independent of the beliefs of speakers, it undermines any argument for thinking that changes in scientific beliefs must change meanings (cf. Putnam 1973).

It is not my intention here to adjudicate between the causal theory and the older idea of theoretical definitions as an account of the semantic workings of scientific (or any other) terms. As it happens, I think there are some good reasons for favoring the old account. To mention just two: (i) the causal theory threatens to ascribe referents to a number of intuitively non-referring terms, such as "phlogiston" (making it refer to deoxygenated gas), "spirit possession" (psychological disturbance), and so on, whereas in reality these terms lack reference; and (ii) the causal theory seems unable to account for terms, like "positron," "neutrino," and "quark," that are explicitly introduced to refer to hypothetical entities which are conjectured to play certain theoretically specified roles, before any direct experimental manifestation of these entities is available for any dubbing ceremony.

Still, as I said, it is not my intention in this paper to argue for the possibility of theoretical definitions and against the causal theory. Rather, I want to address the hypothetical question: *if* some terms have their meanings determined by theoretical definitions, *then* how should we deal with the problems this raises?

I shall proceed as follows. First, in the next section, I shall deal with the worry that theoretical definitions make theories analytic and scientific theory-choice therefore irrational. I shall show that this worry is relatively superficial. A proper understanding of the structure of theoretical definitions will show that the assumptions involved in theoretical definitions have a perfectly good synthetic, empirically assessible content. After this I shall return to the imprecision of theoretical definitions. This is the real philosophical difficulty about theoretical definitions—*which* assumptions play a definitional role? I shall deal with this difficulty in Sections 4–7.

3. The Ramsey-Carnap-Lewis Account of Theoretical Terms. In retrospect, many of the 1960s and 1970s worries about the synthetic status of meaning-defining assumptions can be attributed to excessively verificationist attitudes to meaning, and in particular to the idea that the meaning of a theoretical-defined term is fixed by the observational evidence which warrants its application. The work of Carnap and Quine in the middle decades of the century showed that many terms cannot be given full observational definitions (Carnap 1936, Quine 1951). But even after this, many philosophers continued to equate a term's meaning with the set of paths that lead from observation to its application, and hence to think of a theoretical definition as something which creates a set of such paths (cf. Feyerabend 1962, 1965; Hesse 1974, Papineau 1979).

However, we will do much better to turn our back on verificationism, and ask instead what theoretically defined terms allow us to *say* about the world, that is, about their referential semantics, leaving questions about criteria for application to take care of themselves. If we do this, then worries about the synthetic status of theories and the rationality of science theory-choice will turn out to dissolve themselves.

The key idea needed to understand the referential sematics of theory-dependent terms has long been available. As with so many other problems in contemporary philosophy, Ramsey led the way. His essential insight was to view theoretically defined terms as disguised definite descriptions (Ramsey 1931). This approach was developed further by Carnap (1966), and received its definitive statement in Lewis's "How to Define Theoretical Terms" (1970). On most points in this section I shall follow Lewis.

Suppose that F_1 is a theoretically defined term, and that $T(F_1)$ is the set of assumptions involving F_1 that contribute to its definition. (I here assume that $T(F_1)$ is a precise set, since my aim in this section is merely to show that theoretical definitions do not make defining theories analytic; this assumption of precision will be relaxed in the next section.) As a first approximation, the Ramsey-Carnap-Lewis suggestion is that F_1's meaning is given by the following definition.

$$F_1 = \text{df } (\P x)(T(x)) \qquad \text{(i)}$$

where $T(x)$ is the open sentence that results from $T(F_1)$ when F_1 is replaced by the variable x, and ¶ is the definite description operator.[3]

[3]In the context of theoretical definitions, I prefer to read the definite description operator as yielding genuine singular terms whose function is to introduce entities into our discourse. So I understand $(\P x)(T(x))$ as referring to the unique satisfier of $T(x)$ if there is one, and as failing to refer otherwise. If $(\P x)(T(x))$ lacks reference, I take all atomic sentences involving it to be false, but the negative existential claim $-(Ez)(z = (\P x)(T(x))$ to be true. An alternative view of the definite description operator is that it is a device for abbreviating Russell-style existential quantifications. Nothing much in what follows hangs on the choice between these alternatives, except questions of rigidity in modal contexts. I shall discuss rigidity explicitly in Section 5.

So the Ramsey-Carnap-Lewis idea is simple enough. Theoretical definitions yield terms which refer to whichever entity[4] plays the role specified by T(x), assuming there is one such.

One immediate complication we must deal with is that T(x) will use other non-logical terms apart from F_1 to specify this theoretical role. In many cases these will include further theoretically defined terms $F_2, \ldots, F_n$. Since we are trying to explain the meaning of theoretically defined terms in general, we cannot take these uses of $F_2, \ldots, F_n$ for granted. Nor need we. The solution is to existentially quantify into the positions occupied by these terms, and define F_1 by the equation:

$$F_1 = \text{df } (\P x_1)(E!x_2, \ldots, x_n)(T(x_1, \ldots, x_n)) \qquad \text{(ii)}$$

This says that F_1 refers to the first in the unique sequence of entities which satisfies $T(x_1, \ldots, x_n)$, if there is such a sequence, and fails to refer otherwise (where $T(x_1, \ldots, x_n)$ is the open sentence which results when we replace $F_1, \ldots, F_n$ by $x_1, \ldots, x_n$ in F_1's defining theory.)[5]

Note that some of the non-logical terms involved in defining the Fs had better *not* have their meanings fixed by their theoretical roles, otherwise the necessary existential quantifications will remove all non-logical terms whatsoever from $T(x_1, \ldots, x_n)$ and take away any power it may have to identify a unique set of entities. Traditionally this "mooring" was provided by *observation* terms, with $T(x_1, \ldots, x_n)$ therefore specifying how the theoretical entities relate to each other *and* to observable entities. But I shall not commit myself to the existence of such a class of observation terms, since I need only assume, following Lewis, that $T(x_1, \ldots, x_n)$ is somehow moored by antecedently understood terms whose meanings are independently fixed, not that these antecedently understood terms are necessarily observational.[6]

[4]Since such entities are normally properties (*electrically charged*), kinds (*gas*), and other second-order entities, it is arguable that the quantifiers and definite description operators ranging over them are better represented as second-order than first-order. I myself prefer this approach, and the arguments of this paper would all go through perfectly well if we adopted it. Lewis, on the other hand, reifies the referents of theoretical terms where necessary (*the property of being electrically charged, the gaseous state*) to bring them into the range of first-order quantification. I have followed him in the interests of familiarity and in order to placate those who are suspicious of second-order quantification. Note that on either approach we need to restrict the candidate referents for theoretical terms to *natural* properties or kinds, and exclude gerrymandered or gruesome entities, if we are to capture the intended meaning of theoretical terms.

[5]Do not assume that the same theory will be used in defining all definitionally related Fs. For example, the term "atom" will presumably appear in "electron" 's definition, yet it is arguable that we do not need assumptions about electrons in defining "atom," since we can specify what atoms are without specifying anything about their internal structure.

[6]Indeed it is important, though relatively little remarked, that these antecedently understood terms should not only include predicates (like "square," "red," "adjacent to," and so on), but also terms for the physical links between phenomena so described and theoretical entities (such as "causes," "physically necessitates," or "is probabilistically correlated with").

My main aim in this section is to show that theoretical definitions do not seriously impugn the synthetic status of the defining theories. To see this, suppose that $T(F_1, \ldots, F_n)$ is the theory involved in defining F_1. The problem of synthetic status is supposed to be that this definition of F_1 will turn this theory into an analytic truth. However, it follows immediately from the definition of F_1 given by (ii) and corresponding definitions for the other Fs, that $T(F_1, \ldots, F_n)$ is definitionally equivalent[7] to:

$$(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n)). \qquad \text{(iii)}$$

This claim simply says that there is a unique sequence of entities which bear the relationships to each other and to antecedently identifiable entities specified by T. For any non-trivial T, this will not be a mere matter of meaning. That there should exist the entities required to make (iii) true is a substantial synthetic issue, to be confirmed or disconfirmed by the empirical evidence.

To take an example, we might take "atom" to be defined via the assumptions that: (a) atoms are the smallest parts of matter separable by chemical means, (b) there is a different species of atom for each element, and (c) atoms of different species combine in simple whole number ratios. Under the Ramsey-Carnap-Lewis treatment, the conjunction of these defining assumptions is equivalent to the claim *that there are entities which (a') are the smallest parts of matter separable by chemical means (b') are different for different elements, and that (c') combine in simple whole number ratios determined by their elements.* The assumption that (a)–(c) are definitional of "atom" clearly does not imply that this italicized claim is an analytic truth. For it is perfectly possible that the smallest units of chemically separable matter do not form a different species for each element, and even if they do, that they do not combine in simple whole number ratios. It is an empirical discovery that there are entities of which all these things are true—namely, the discovery that there are atoms.

This last point illustrates a general feature of the Ramsey-Carnap-Lewis approach to theoretical terms. Definitions like (ii) mean that we can *eliminate* theoretically defined terms from any claims in which they appear. Thus suppose $S(F_1)$ is any claim involving theoretical term F_1, and that $T(F_1, \ldots, F_n)$ is the theory that defines F_1 (with $F_2, \ldots, F_n$ the other

This kind of terminology is necessary because any attempt to specify the relevant links between the intended theoretical definienda and antecedently identifiable phenomena using universal quantification alone will inevitably allow unintended interpretations.

[7]Note that on the reading of the definite description operator I prefer this will not be a *logical* equivalence (since a given entity's satisfying T(x) will not logically require that only one entity does). But, even so, it will be a consequence of F's definitional equation with a definite description so understood that $T(F_1, \ldots, F_n)$ and $(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n))$ must have the same truth value.

theoretical terms used in the definition). Then $S(F_1)$ will be definitionally equivalent to:

$$(E!x_1, \ldots, x_n)((T(x_1, \ldots, x_n) \,\&\, S(x_1)), \qquad \text{(iv)}$$

that is, to the claim that there is a unique sequence of entities which satisfy $T(x_1, \ldots, x_n)$ and the first element of this sequence also satisfies S(x). This claim says what $S(F_1)$ says, but without using the term F_1.[8]

This eliminability of theoretical terms points to an important moral. Namely, that the use of theoretical terms defined in the Ramsey-Carnap-Lewis way cannot give rise to any serious philosophical problems (assuming still that the definitions are precise). For any claims formulated using such terms are simply a shorthand for claims that can be formulated without such terms, by instead existentially quantifying into the places those terms occupy. The adoption of a shorthand can scarcely itself be responsible for substantial philosophical difficulties.[9]

The reason we often need this shorthand is that the equivalent claims which eliminate the shorthand will generally be much more complicated to articulate. Thus it is much easier to say that "There are two atoms in molecules of hydrogen gas" than "There exist entities which are the smallest chemically separable parts of matter, one species for each element, which combine in small whole number ratios, and hydrogen molecules contain two of them." And the longhand version would be even more complicated if we existentially quantified into the relevant assumptions about "molecule" as well, not to mention "hydrogen" and "element."

Note that once we do adopt the convenience of shorthand theoretical terms, then this will yield *some* new analytic claims involving them, namely any claims which follow from definitions like (ii). But it would be a con-

[8]Again, the quantified claim (iv) will not express quite the same proposition as $S(F_1)$ on my preferred reading of the definite description operator. But it will be a consequence of F_1's definition that (iv) and $S(F_1)$ must have the same truth value.

[9]Thus consider "the Duhem-Quine thesis," according to which decisions on specific hypotheses can never be separated from decisions on some larger theoretical framework. There may be some good arguments for some versions of this kind of "confirmational holism." But theoretical definability does not in itself provide such an argument. True, theoretical definitions are in a sense holist, involving all the assumptions in the relevant T(F), and this does mean, as I shall point out below, that any claim made using a theoretically defined F commits us to *all* of the defining T. But to infer that there is therefore no question of adjudicating the parts of T separately is to assume that the term F provides the *only* way of stating what the theory claims. However, we can always avoid F and switch to explicit quantification into T(F). The advantage of this move, in the present context, is that it allows us to break T down into various weaker claims, which can be assessed against the evidence separately. (Thus we could consider (E!x)(T*(x)) for various weaker T*s which follow from the total T; we could also consider some corresponding existential claims without uniqueness.) Now it may be that various of these claims would be confirmationally bound up with each other, even after we had performed the decomposition. But this is a matter for detailed investigation, and certainly is not an immediate consequence of the fact that *one* way of talking about the entities postulated by these claims commits us to the conjunction of these claims.

fusion to think that this somehow illegitimately turns the original factual content of synthetic theories into definitional truths. The factual content of any such defining theory T is still given by a Ramsey-style sentence of the form

$$(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n)). \qquad \text{(iii)}$$

The further analytic truths that are introduced along with shorthand Fs, simply specify, so to speak, what the entities required by (iii) are *called.* For instance, the analytic truth

$$(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n)) \supset (E!x_2, \ldots, x_n)T(F_1, x_2, \ldots, x_n)$$

says that *if* there is a unique sequence of entities which satisfies T, *then* the first member is F_1.

There is one obvious disadvantage to theoretically defined Fs which function as a shorthand for complex Ramsified sentences in the way I have outlined. Since the Ramsified sentences in question are all prefixed by the claim $(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n))$, it will not be possible to use such an F to make positive assertions which deny any part of this claim. In contexts where we have reason to withdraw an assumption to which we were once committed and in terms of which we have introduced the Fs, this may well be an inconvenience. But it is important to recognize that it is *only* an inconvenience, not something which forces us into incoherence or which threatens the objectivity of science. For we can always revert to an F-free formulation to say what we want to say. For instance, if we no longer believe $(E!x_1, \ldots, x_n)(T(x_1, \ldots, x_n))$, but only $(E!x_1, \ldots, x_n)(T'(x_1, \ldots, x_n))$ where T′ omits or varies some of the assumptions in T, then we can say what we want to say in just this way, and use sentences which quantify explicitly into $T'(x_1, \ldots, x_n)$. And if we want a shorthand which frees us from the need to spell out T′ every time we say anything, we can define a new set of theoretical Gs, say, to serve this purpose.

Note that it would be wrong to assume, as much of the "meaning-variance" literature does, that when such a theoretical change takes place, and we stop believing F_1's defining theory $T(F_1, \ldots, F_n)$, that the meaning of F_1 will change therewith. For F_1 can still have the same definition as it always had, specifying that it refers to the first element in the unique satisfier of $T(x_1, \ldots, x_n)$, if there is one such, and to nothing otherwise. Indeed, even after we stop believing $T(F_1, \ldots, F_n)$ we will still sometimes have occasion to use F_1 in just this sense, namely, when we deny that F_1s exist.

The point is that you don't have to *believe* $T(F_1, \ldots, F_n)$ in order to accept the corresponding definition of F_1 and use F_1 accordingly. That is, you can agree that F_1 refers to the unique satisfier of $T(x_1, \ldots, x_n)$, if there is one such, and to nothing otherwise, quite independently of whether you

believe there is such a unique satisfier. And so if you start off believing $T(F_1, \ldots, F_n)$, but then stop doing so, this does not stop you continuing to use F_1 with the same meaning.

Recall my earlier remarks about excessive verificationism. On the verificationist picture, the meaning you attach to a theoretical term F can be equated with your set of dispositions to apply it or its negation in response to sensory evidence. Meaning as so conceived will depend on what theory about Fs you *believe,* since this will affect what sensory evidence you take to indicate the presence or absence of Fs. Correspondingly, if you change your theory, you will change the meanings of your theoretical terms.

There is no such implication on the Ramsey-Carnap-Lewis account of theoretical terms. When you change your theory, and cease to believe in the entities to which your old theory committed you, this will indeed change your dispositions to apply its terms in response to sensory evidence. You will no longer say, for instance, that the combustion stops because "the air is saturated with phlogiston." But this will not be because you have changed the meanings of your terms, but simply because you have new beliefs about the connection between observational phenomena and facts about phlogiston.

In making this point, I do not of course want to deny that meaning changes are *possible.* Scientific and other theoretically defined terms are of course capable of shifts in meaning, just like any other words, and such shifts raise interesting issues. My point is merely that there is nothing in the structure of theoretical definitions as such to imply that the meaning of defined terms *must* shift whenever belief in the relevant theory changes. The view that such shifts are inevitable is nothing but a vestigial hangover from the verificationist model of meaning.

4. Imprecision without Tears. I turn now to the question of how *much* theory contributes to the definitions of theoretically defined terms. In addressing this question, I shall take it to be a basic desideratum on such terms that they be useful for stating truths (and here I mean truths other than such negative existentials as "There are no Fs"). Given this, we can note there are two dangers a definition of a theoretical term F must avoid. Remember that any (non-negative-existential) claim made using F implicitly asserts that there is a unique satisfier for the defining open sentence $T(x_1, \ldots, x_n)$. So a first desideratum is that the definition must include *enough* theory to ensure that T is *uniquely* satisfied. If so little is included in T that a number of different entities satisfy the requirements imposed by T, then every (non-negative-existential) F-claim will fail for this reason. Second, the definition must *not* include *so much* theory that *nothing* satisfies T. If so much is included in T that no entities satisfy all the require-

ments imposed by T, this too will falsify every (non-negative-existential) F-claim.

It is central to what follows that the satisfaction of these two desiderata does *not* require that F have a precise definition. In this section and the next, I shall illustrate this point with a simplified model of imprecise definitions. I shall make the story more realistic in Section 6.

Consider a case where certain core assumptions involving F (T_y – "y" for "yes") unquestionably *do* contribute to F's definition, and other accepted assumptions involving F (T_n – "no") unquestionably do *not* contribute, but that beyond that it is indeterminate whether any other generally accepted claims involving F (T_p – "perhaps") have a definitional status. As long as T_y is satisfied by only one entity, then enough is unquestionably *in* the definition to ensure a unique satisfier. And as long as T_y *plus* T_p is also satisfied by that entity, then enough is unquestionably *out* to ensure that the definition does not lack satisfiers altogether.

The central point I want to make in this paper is that if an F has this kind of imprecise definition, with T_y strong enough to ensure a unique referent and T_y-plus-T_p not too strong to rule out a referent entirely, then the imprecision does not matter. For note that in these cases *F would end up referring to the same entity however the imprecision were resolved.* Given this, there is no need to resolve the imprecision. Understand F as you will, consistently with your definition including T_y and excluding T_n, and you will be referring to the same thing. I would say that in this kind of case F has a definite reference, even if it is indeterminate exactly which assumptions involving F fix this reference.

Not all terms with this kind of imprecise definition will have this attractive feature. Sometimes T_y will fail to ensure a unique satisfier, or T_y-plus-T_p will be so strong as to rule out satisfiers altogether. But let us take cases like this to be the exception rather than the rule, and postpone discussion of them until Section 7. For the moment I want to focus on the more benign kind of case, where there is determinate reference despite imprecision of definition.

In a sense, theoretical terms with the above kind of imprecise definitions will be vague. But the kind of vagueness involved is not straightforward. It is certainly vague which assumptions involving F should be taken as constituting its Ramsey-Carnap-Lewis definition. This is just another way of formulating the model of theoretical definitions I have just proposed. But since vagueness in this sense does not necessarily imply any indeterminacy of reference, it does not automatically follow that any sentences involving F will lack determinate truth conditions. If the referent of F is determinate, then the condition which has to be satisfied for any sentence S(F) to be true will be determinate too (unless of course there is indeterminacy in the referential values of *other* terms in S(F)).

This is why imprecision in theoretical definition need not matter. I take it that it is preferable, for scientific purposes anyway, that all our claims should have determinate truth conditions. Definitional imprecision that threatens truth conditional determinacy is therefore to be avoided. But there is nothing obviously worrying about definitional imprecision that leaves all truth conditions determinate.

5. Modal Contexts. In arguing that definitional imprecision is consistent with determinate truth conditions. I have implicitly been ignoring modal constructions. Modal sentences raise further issues.

The interpretation of modal sentences involving theoretically defined terms hinges on whether those terms are rigid. As it happens, I do think that theoretically defined terms are rigid, because of their equivalence to definite descriptions, together with my understanding of the definite description operator as forming singular terms whose function is to introduce entities to our discourse. However, a number of delicate issues are involved here, and I shall not premise my discussion on this. Rather, I shall show that modal sentences involving theoretically defined terms raise no substantial extra problems, whether or not those terms are rigid.

Suppose first that some theoretically defined term F is rigid, that is, it refers in modal contexts to whatever it refers to in the actual world. Since we are assuming that F has a determinate referent in the actual world, this means that modal claims involving such a rigid F will have determinate truth conditions too—unless, of course, there is referential indeterminacy in other parts of the modal claim.

In fact, this last caveat about vagueness arising from *other* components in modal F-claims is all too likely to apply. Consider this sentence: "Atoms necessarily have nuclei." Assume that "atom" is rigid and this sentence therefore attributes a *de re* necessity. My own reaction is that, even on this assumption, this sentence is indeterminate in truth value. Nothing decides whether particles otherwise like atoms but internally homogeneous are *atoms* or not.

However, the indeterminacy here is not due to any indeterminacy in the referential value of "atom." If "atom" is rigid, we are talking determinately about our atoms, the things which play the atom role in our world, and which undoubtedly do have nuclei. The question is whether nuclei are essential or accidental to these things. If the answer is indeterminate, this is not because "atom" has indeterminate reference, but rather because of imprecision in such notions as essential property or *de re* necessity.

Now assume that theoretically defined terms are not rigid, in the sense that when they occur inside modal operators, they refer variably to whatever satisfies their definition in different counterfactual circumstances, rather than to what satisfies their definition in the actual world. Of course,

terms like these can still occur outside modal operators and thus feature in *de re* modal claims, and in such cases the points just made in connection with rigid terms will apply. But when non-rigid theoretically defined terms occur within modal contexts, then any imprecision in their definitions can lead to indeterminacy of truth value.

To have an example, let us assume that the three assumptions specified earlier (chemical indivisibility, different atoms for different elements, simple whole number combinations) are definitely part of the definition of "atom" (that is, in "atom" 's T_y), but that it is undecided whether the assumption that atoms have nuclei is part of "atom" 's definition (so this assumption is in "atom" 's T_p). Now consider the claim "Necessarily, atoms have nuclei." If "necessary" has wide scope here and "atom" refers variably in modal contexts, then this claim will have an indeterminate truth value. For take a possible world in which there are things which satisfy T_y (that is, are the smallest chemically separable parts of matter, etc.) but do not have nuclei. Does "atom" refer to these things at this possible world? If the assumption that atoms have nuclei were criterial for "atom," then the answer would be "no;" if it were not, the answer would be "yes." But since the assumption that atoms have nuclei is in "atom" 's T_p, the answer is indeterminate. So it is indeterminate whether or not these are worlds in which atoms lack nuclei, and in consequence it is indeterminate whether "Necessarily, atoms have nuclei" is true.

Is it an objection to imprecise definitions of theoretical terms that, when these terms are treated non-rigidly, we get modal claims which lack determinate truth conditions? It is obviously no objection if, like me, you take theoretical terms to be rigid. But, even if you do not want to commit yourself to rigidity, this kind of modal indeterminacy ought to be unworrying. For, as far as I can see, it is unimportant to science, or anything else, to resolve such questions as whether it is necessarily true that atoms have nuclei. That it is *true* that atoms have nuclei is of course a matter of great significance, but nothing seems to hang on whether this truth is in addition *necessary.* So it will not matter if such claims of necessity lack definite truth values.

It is important that I am talking here about necessity *tout court,* not about "physical necessity." Questions of physical necessity are, of course, central to science, and indeed scientific theories can be read simply as delineations of what is physically necessary. But discriminating the physically necessary does not require the finer discrimination of the absolutely necessary. After all, anything which follows from true scientific theory is physically necessary, and this will unequivocally include "atoms have nuclei," along with all other claims that follow from T_y-plus-T_p.[10] However,

[10]Remember that I am assuming in this section that both T_y and T_y-plus-T_p pick out a unique satisfier. This means that T_y-plus-T_p comes out true on all of the candidate definitions for F.

once science has delineated the physically necessary in this way, there seems no obvious reason why it should concern itself in addition with what is and is not absolutely necessary.

6. More Realistic Imprecision. How different is the T_y-T_p-T_n model assumed in the last two sections from the traditional analytic-synthetic distinction? The obvious difference is that this model makes the criterial status of the assumptions in T_p indeterminate. Nevertheless, some readers will no doubt feel that this is matter of relative detail, and that many of the objections to a simple analytic-synthetic distinction apply equally to my tripartite division of theoretical assumptions into analytic, synthetic, and indeterminate.[11]

This is not the place for a full discussion of the analytic-synthetic division. My own view is that the issue is an essentially empirical matter about the use of the relevant terms by the relevant linguistic community. Accordingly, I see no reason in principle why evidence should not show that a given community treats the various assumptions involving some term F in line with the T_y-T_p-T_n model.[12]

At the same time, I suspect that even this triparite division offers far too simple a picture of the actual use of most theoretically defined terms.

[11]In fact this is not quite my division. As the discussion in section 3 showed, a theoretical definition does not make the defining T(F) itself analytic, but only the "Carnap sentence" $(E!x)(T(x)) \supset T(F)$.

[12]In order to find out whether this model applied to a given linguistic community, we could ask the members of the community what they would say if it turned out that, contrary to their opinion, that there were no unique satisfiers of (a) T_y, or (b) T_y-and-T_p, or (c) T_y-and-T_p-and-T_n. If their answers were (a) this would show that there are no *F*s; (b) that they are unusure what this would imply for the existence of *F*s; and (c) that this would simply show that T_n is false of *F*s, not that there are no *F*s, then we would have good evidence that they conform to the proposed model. Note that it matters to our appreciation of this evidence that we think of theoretical definitions in the Ramsey-Carnap-Lewis way, rather than on the verificationist model in which theories affect meanings by influencing dispositions to respond to sensory evidence. On the verificationist picture, facts about what users would say if some assumption turned out false cannot provide decisive evidence that this assumption is critical for F. For, on the verificationist model, the rejection of a criterial assumption for F would require users to change F's meaning *somehow,* since they would therewith stop applying the term as advised by that assumption; however, they could either change meanings so as to stop applying the term altogether ("there are no Fs"), or so as to apply it on the basis of their remaining assumptions alone ("there are Fs"). So even if users would continue to say "there are Fs," were they to reject some assumption A, this does not, on the verificationist model, show A is not currently criterial for F. On the Ramsey-Carnap-Lewis view, by contrast, rejecting a meaning-constituting assumption imposes no pressure at all to change meanings (see the end of Section 3). So if users are sure they would continue to say "there are Fs," if they were to reject some assumption A, the obvious inference to draw is that A is not criterial for F. (True, this inference is defeasible. Even on the Ramsey-Lewis-Carnap view, users *could* change the meaning of F, were they to abandon A, and so continue to say "there are Fs," even though A is currently criterial for F. But this diagnosis of why users feel sure they would retain "there are Fs," if they were to abandon A, would surely require some extra evidence, such as their announced intention to change meanings in that case, in order to defeat the obvious explanation that A is not at present criterial for F.)

Accordingly, in this section I shall explore some less simple models. I do not intend this to be an exhaustive survey of the possibilities. My main aim in this section is simply to show that, even if more complex models of theoretical terms prove necessary, this need not undermine my central point that determinacy of truth conditions can coexist with imprecise definitions.

One obvious respect in which we can expect the T_y-T_p-T_n model to prove overly simple is in respect of the sharpness of the T_y-T_p and T_p-T_n divisions. This is a general point about vagueness: when there is a penumbra of vagueness between "yes" and "no" answers to some question, it will normally also be vague when "yes" shades into "perhaps" and "perhaps" into "no." Similarly, in the present context we can expect that some theoretical assumptions will be indeterminately located with respect to the T_y-T_p or T_p-T_n borderlines. However, this need not make any difference to the possibility of determinate reference and truth conditions. For the referent of F will still be determinate, as long the accepted assumptions definitely in T_y (which are not T_y-T_p borderline) still suffice to pick out a unique referent, and as long as this referent still satisfies our definition if we leave out the assumptions definitely in T_n (though not necessarily those on the T_p-T_n-borderline).

David Lewis says in passing in his "Psychophysical and Theoretical Identifications" (1972) that we should not define a term by the whole of the accepted theory involving it, since this would mean the term will have no referent if any part of that theory is false. Instead he suggests that such a term should be defined as referring to that unique entity, if any, which satisfies the disjunction of all conjunctions of *most* theoretical assumptions involving it. This suggestion has some similarity to the approach adopted here, in that it seeks a definition which is simultaneously weak enough to allow satisfaction while strong enough to ensure uniqueness. But there are two important differences between Lewis's suggestion and mine. First, he accords all assumptions involving a term equal weight in defining it, and has nothing corresponding to my T_y-T_p-T_n classification. Second, it is not clear that Lewis has in mind an imprecise definition, since the "disjunction of all conjunctions of most assumptions" implies a quite definite condition, if we take "most" at face value as simply meaning "more than half."

Lewis's idea of a disjunctive definition is attractive. It is interesting to consider how it might be detached from (a) his ascription of equal definitional weight to all assumptions and (b) the definiteness which attaches to his use of "most." Take point (b) first. It seems wrong to invoke precisely those conjunctions containing more than *half* the relevant assumptions; and in any case there are well-known difficulties about counting theoretical assumptions. A more attractive strategy would be to frame the definition in terms of the disjunction of all conjunctions of *large enough*

sets of assumptions. What counts as "large enough" will of course be vague. But, in line with my overall approach, this won't matter, as long as at least one conjunction is satisfied even when we set "large enough" high, and as long as no conjunction is multiply satisfied even when we set "large enough" low.

The other objection to Lewis's suggestion was that it counts all assumptions as of equal definitional importance. I avoided this in the previous two sections by postulating the T_y-T_p-T_n division. Perhaps an even better approach would be to treat the accepted assumptions involving a term F as forming some kind of order in respect of their definitional importance. We could then say that an entity qualified as the referent of F if it achieved a reasonably high "score" in satisfying these assumptions, with "more definitional" assumptions contributing more to this score than "less definitional" ones. Note that this would not only give different definitional weight to different assumptions, but would also incorporate the disjunctive idea, in that an appropriate score might be compiled in different ways, and also the idea of imprecision, in that what counts as "reasonably high" will be vague.

This last suggestion could obviously be refined in various ways, and one can imagine further analogous models. But I shall not continue this discussion any further, since the central point I want to make in this paper will apply to *any* account of theoretical definitions which appeals to an imprecise condition which is vague between a minimal version which is relatively easy to satisfy and a maximal version which is relatively hard to satisfy. My central point, to repeat, is that this kind of definitional imprecision will not give rise to indeterminate truth conditions as long as only one entity satisfies the minimal version of the definition and at least one satisfies the maximal version.[13]

Earlier I referred briefly to the contemporary debate about "semantic holism." The possibility that theoretical definitions may be imprecise puts this issue in a rather new light. If every theoretical assumption must either be included in or excluded from a given term's definition, then the fact we cannot distinguish which assumptions are which seems to leave us with no options apart from rejecting theoretical definitions altogether, or allowing that all accepted assumptions contribute equally to such defini-

[13]What if different *individuals* in a given linguistic community give different theoretical definitions to some term, perhaps attaching different definitional weights to certain assumptions, or setting the minimal and maximal limits of the definition's imprecision in different places? Well, they would still all be referring to the same entity, as long as that entity was uniquely picked out by each person's minimal version, and that entity still satisfied each person's maximal version. There are, of course, general questions about how far such a common reference is sufficient to ensure that a given term is univocal in different mouths, even though its users attach different "concepts" to the term. But these questions are not peculiar to imprecisely defined terms.

tions. But if theoretical definitions can be imprecise, there are further alternatives. On the T_y-T_p-T_n model, there is no sharp line between criterial and non-criterial assumptions, yet T_n assumptions definitely do not contribute to definitions. Similarly, on the more flexible kind of "definitional ranking" model outlined in this section, we can expect the majority of non-central assumptions involving a term to have a minimal definitional ranking, and so never to play any definitional role. Quine's most persuasive insight was always that there is no fact of the matter, for many theoretical assumptions, whether they are definitional or not. The idea of imprecise theoretical definitions accommodates this insight, without driving us to the unacceptable conclusion that all accepted assumptions involving a theoretically defined term are criterial for its application.

7. Necessary Refinements. My overall argument has been that even if a theoretical definition is imprecise, the defined term can still have a determinate referent. However, definitional imprecision will not always have this happy outcome. Earlier I observed that, even if an imprecisely defined term has a determinate reference in this world, it will have indeterminate reference at certain other possible worlds (if treated as non-rigid). I argued that this need be no demerit in a theoretical definition. But it would certainly be a demerit if the actual world turns out to be a world in which the defined terms lack a determinate referent.

My full view is that imprecise theoretical definitions are *usually* benign, in that they do not *usually* lead to indeterminacy of truth conditions, and that the points made in the last section therefore show it would be inappropriate to seek to eliminate such definitional imprecision whenever it occurs. But at the same time I admit that there are also cases where such definitional imprecision *does* lead to claims which lack determinate truth-conditional content. When this happens, our discourse is flawed. So when we identify such cases, we ought to remedy the imprecision.

It will simplify the remaining exposition without loss of generality if we let the T_y-T_p-T_n model stand for the general idea of minimal and maximal versions of an imprecise definition. Recall that there are two dangers a theoretical definition of some term F must avoid, if it is to yield a term that is useful for stating truths: it must not make the definition so weak as to fail to identify a unique satisfier; and it must not make the definition so strong that it rules out satisfiers altogether. The special risk facing imprecise definitions is that indeterminate status of the assumptions in T_p can make it indeterminate whether these two dangers have been avoided. Thus, to take the first danger, T_y might be too weak to ensure a unique satisfier by itself, but may be able to do so if conjoined with some of the assumptions in T_p. Then it will be indeterminate whether F refers uniquely. Second, it may be that nothing satisfies all of T_y-plus-T_p, but that there

would be satisfiers if we dropped some of the T_p assumptions from this conjunction. Then it will be indeterminate whether F refers at all.

Cases of the second kind are perhaps more familiar. The example of *belief* from Section 2 fits this bill. Suppose that the assumption that beliefs have internal causal structure is in the T_p of "belief" 's definition. And suppose further that the connectionists are right to deny that the entities which satisfy the undisputed criteria for being beliefs ("belief" 's T_y) have no internal causal structure. (I do not think they are right, but shall assume so for the sake of the argument.) Then it would be indeterminate whether "belief" refers to those entities or not, since it would be indeterminate whether their lack of internal causal structure disqualifies these entities from being beliefs. It is plausible that many cases from the history of science and elsewhere have the same structure. Does the failure of "caloric is a fluid" mean there is no caloric? Does the failure of "straight lines are Euclidean" mean there are no straight lines? Does the failure of "witches have magic powers" mean there are no witches? Does the failure of "entropy invariably increases in a closed system" mean there is no entropy? And so on.

There are also plausible cases of the other kind, where the indeterminacy of T_p makes it indeterminate which entity, if any, some term refers to. Thus modern microbiology tells us that various kinds of chunks of DNA satisfy the undisputed criteria for "gene," and that further assumptions are needed to narrow the referent down. Similarly, relativity shows that both rest mass and relativistic mass satisfy the original Newtonian definition of mass as proportional to amount of matter, and that further criteria are required to render the referent of "mass" unique.

In both these kinds of cases, some new discovery makes it manifest that the looseness in the definition of some term F is not benign after all. Previously we did not worry about exactly what was required to be an *F*, because it seemed not to matter. So, for example, we did not worry whether straight lines needed to be Euclidean, since we took it for granted that anything satisfying the basic criterion for a straight line (by being the shortest distance between two points) would also be Euclidean. But now we realize that we were wrong, and that this definitional imprecision means that nothing decides which of "there are no straight lines in physical space" or "there are physically straight lines, but they are not Euclidean" is true.

The obvious remedy in this kind of situation is to refine the definition so as to resolve the question. We can include the assumption of Euclideanism in the definition of straight line (which will make "there are no straight lines" true), or we can exclude it (which will make "there are straight lines, but they are not Euclidean" true).

Is there anything which makes it right to go one way rather than another

in such cases? I doubt it. Certainly if we look at the history of science, there is no obvious principle which decides whether in such cases scientists conclude that *F*s do not exist, or, alternatively, that there are *F*s, but the assumption at issue is false of them. Consider the different fates of the terms "caloric" and "electricity." Originally both of these were taken to refer to a fluid that flowed between bodies (from hot to cold bodies in the one case, from positively to negatively charged bodies in the other). Later it was discovered that neither quantity is a fluid, and that the appearance of flow is in both cases a kinetic effect. The two cases are structurally similar. Yet we now say that electricity exists, but that caloric fluid does not.

If there is a pattern governing which way the terminology goes in such cases, it is probably one involving the micro-sociology of the thinkers responsible for the relevant theoretical revision, rather than any substantial semantic or empirical facts. Theorists who want to present themselves as merely continuing the tradition of those who have previously studied *F* will retain the term F for the thing satisfying the basic criteria T_y but not the newly revised part of T_p. On the other hand, theorists who want to distance themselves from the existing theoretical establishment will urge that *F* does not exist, and that their new assumption identifies a hitherto unknown entity *G*. It is not at all implausible that the balance of these kind of conservative and radical tendencies among theoretical innovators led to the retention of terms like "electricity," "straight line," and "entropy," on the one hand, and to the abandonment of "witch" and "caloric," on the other.

The idea that scientific decisions are inevitably determined by sociological factors seems antithetical to any realist attitude to science. If mere self-interest and positioning in scientific politics determines what scientists say, and the evidence plays no part, then there seems little chance that the scientists will end up with the truth.

However, the limited role I have ascribed to sociological factors has no such implication. For I am suggesting only that sociological factors come into play when a scientific term that was hitherto thought to have a determinate reference turns out to be vague in a way that requires remedying. When the meaning of a vague term needs refining, there will generally be no objective reason to refine it in one direction rather than another. It is simply a matter of deciding how to use words, given that our previous practice with these words has proved inadequate. So the intrusion of sociological factors at this point need cause no disquiet to the realist.[14]

[14]Stich's most recent remarks (1993) on theoretical definitions agree with me that new decisions on how to use defined terms are often determined by political maneuverings among scientists. However, he does not offer any explicit account of the semantics of theoretical terms (beyond retracting his earlier view that word-world reference relations are unimpor-

REFERENCES

Carnap, R. (1936). "Testability and Meaning", *Philosophy of Science 3:*419–471, *Philosophy of Science 4:*1–40.

———. (1966), *Philosophical Foundations of Physics.* New York: Basic Books.

Devitt, M. (forthcoming), *Coming to our Senses: A Naturalistic Program for Semantic Localism.* Cambridge: Cambridge University Press.

Feyerabend, P. (1962), "Explanation, Reduction and Empiricism", in H. Feigl, G. Maxwell, and M. Scriven (eds.) *Minnesota Studies in the Philosophy of Science 2.* Minneapolis: University of Minnesota Press, pp. 231–272.

———. (1965), "On the 'Meaning' of Scientific Terms", *Journal of Philosophy 62:*266–274.

Fodor, J. (1987), *Psychosemantics.* Cambridge, MA: MIT Press.

Hesse, M. (1974), *The Structure of Scientific Inference.* London: Macmillan.

Kuhn, T. S. (1962) *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

Lewis, D. (1970), "How to Define Theoretical Terms", *Journal of Philosophy 67:*427–446.

———. (1972), "Psychophysical and Theoretical Identifications", *Australasian Journal of Philosophy 50:*249–258.

Papineau, D. (1979), *Theory and Meaning.* Oxford: Clarendon Press.

Putnam, (1973), "Explanation and Reference", in G. Pearce and P. Maynard (eds.), *Conceptual Change.* Dordrecht: Reidel, pp. 199–221.

Quine, W. V. O. (1951), "Two Dogmas of Empiricism", *Philosophical Review 60:*20–43.

Ramsey, F. (1931), "Theories", in his *The Foundations of Mathematics.* London: Routledge and Kegan Paul.

Ramsey, W., S. Stich, and J. Garon, (1990), "Connectionism, Eliminativism and the Future of Folk Psychology", in Tomberlin, J. (ed.), *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind:*499–533.

Scheffler, I. (1967), *Science and Subjectivity.* New York: Bobbs-Merrill.

Shapere, D. (1966), "Meaning and Scientific Change", in R. Colodny (ed.), *Mind and Cosmos.* Pittsburgh: University of Pittsburgh Press, pp. 41–85.

Stich, S. (1991), "Do True Believers Exist?", *Arisotelian Society Supplementary Volume LXV:*229–244.

———. (1993), "Concepts, Meaning, Reference and Ontology: A Reply to Frank Jackson", in K. Neander and I. Ravenscroft (eds.), *Prospects for Intentionality, Working Papers in Philosophy 3,* produced by the Research School of Social Sciences, ANU, Canberra, pp. 61–77.

tant, since semantic descent would then imply that all claims are unimportant), and so at first sight seems to allow the non-realist conclusion that sociological considerations often decide substantive scientific issues.