Meaning and the Fixation of Belief

Hilary PUTNAM
Harvard University

If one insight is characteristic of twentieth century approaches to meaning and reference, it is that these problems cannot be discussed in isolation from the understanding of belief fixation as a whole. Calling this insight «twentieth century» is not to claim temporal priority for our contemporaries: Kant certainly saw these problems as deeply interconnected, and so did William James, much of whose work appeared in the nineteenth century. But the extent of the interconnection has been emphasized an unprecedented degree by the twentieth century American philosopher Willard Van Quine, and his ideas are today even more central to the discussion than they were when his famous article, «Two Dogmas of Empiricism», appeared in 1950.

Traditionally, rational belief fixation is thought of as a topic to be subdivided between deductive logic and a somewhat tenouous subject optimistically called «inductive logic». (In addition, C. S. Peirce, the originator of pragmatism, separated theory construction from inductive logic under the name «abduction».) Quine does not assume that inductive logic or Peirce's «abductive inference» really exist as definite methods that can be formalized, what he emphasizes is the informality of the procedures by which choose in different contexts between preserving what think are our observations and preserving the theories or theoretical principles that we cherish. The central ideas are that «revision can strike anywhere» —a fallibilism that puts Quine squarely in the tradition of Peirce and Dewey— and that one's beliefs on any topic may become relevant to the fixation of beliefs on any other topic. (Illustrations from the history of science are easy to give: consider how unexpected darkening of a photographic plate in the laboratory of Pierre Curie turned out to be relevant to the possibility of an atomic bomb, or —after the existence

Teorema, n.º 15/3-4 - 1986-87-88-89 - 127-146. Editorial Universidad Complutense. Madrid

and nature of radioactivity had been worked out, and applied to determining the age of organic and inorganic materials via the technique of carbon dating— to the age of humanoid fossils.) The idea that our beliefs are all potentially interconnectible and that revision can strike at any point of the theoretical network is often referred to as «holism», and Quine may be called the outstanding contemporary advocate of holistic theories both of belief fixation and of meaning.

At the opposite extreme from Quine's holism are the views associated with Logical Positivism, which held that each empirically meaningful sentence had its own range of confirming and disconforming experiences (or, in a famous slogan, its own «method of verification»), and that this was fixed by the «empirical meanings» of the terms in the sentence.

The position just described as the «opposite extreme» from Quine's holism is one which is committed to the claims that (1) one could in principle test a scientific theory *one sentence at a time;* and (2) that this is so because each empirical sentence contains terms with a certain kind of *meaning.* Positivist reductionism is a position about belief fixation and simultaneously a position about meaning.

Rudolf Carnap's celebrated book *The Logical Construction of the World (Der Logische Aufbau der Welt)* held that every scientific term should, in principle, be definable from an «autopsychological basis»: from terms referring to one's own sensations. (So each scientist could interpret every scientific assertion as an assertion about his or her private experiences.) It was in reaction to this position of Carnap's, in large part, that Quine was inspired to put forward holistic views of both belief fixation and meaning.

Today there are few, if any, philosophers who would not agree that belief fixation *in the mature sciences* is holistic. Several sorts of consideration have played a role in bringing about this unusual unanimity. First of all, it is generally recognized that scientific testing depends very critically on the use of *auxiliary theories and hypotheses,* including idealizations of every kind. Newton's theory of gravitation, by itself, is compatible with any orbits whatsoever, for example. We can even reconcile Newton's theory with square orbits by saying, «non-gravitational forces must be acting on the system». Only in the presence of auxiliary hypotheses (ones which are often implicit in the physicist's calculations, rather than being explicitly stated as part of the theory), such as «only gravitational forces are acting on the system», «the only significant gravitational sources are the sun, earth...», etc., can we derive orbits or other testable predictions from Newton's theory. In this sense, it is not only the case that we cannot speak of «the empirical meaning» of the Law of Universal Gravitation by itself; we cannot even speak of «the empirical meaning» of *Newtonian mechanics* as a whole, unless we resort to the artificial expedient of taking the currently accepted auxiliary hypotheses to be a part or «Newtonian mechanics». (What «Newtonian mechanics» *is* changes with each pro-

blem the physicist approaches, if we adopt this expedient; and it is still true that «empirical meaning» belongs to the whole system under test in each case, and cannot be ascribed to any one sentence in isolation.) Secondly, even when a scientific term possesses a so-called «definition», that definition has no special epistemic privilege. For example, even if the sentence «momentum is mass times velocity» was a «definition» two or more centuries ago, while the sentence «momentum is conserved» originally expressed an «empirical consequence», scientists in the present century did not regard it as more «sure» that momentum is mass times velocity (let alone regard it as an «analytic truth») than Special Relativity. In fact, Einstein showed (assuming Special Relativity) that if there *is* a conserved vector quantity (in the direction of motion of the particle), which is approximately equal to mass times velocity in the limit of small velocities, then it *cannot* be exactly equal to mass times velocity. The conclusion he drew was that *it turns out not to be strictly speaking true* that momentum is mass times velocity. The sentence that was *originally* a «definition» was *overthrown.*

The principle that played the crucial role in overthrowing $p = mv$ (momentum is mass times velocity) was the Principle of Special Relativity. $p = mv$ wasn't overthrown, as we might say, *directly* —by a clash with an «observational consequence»— but by a clash with a well-supported principle that scientists liked better. It was the atractiveness of one construction of physical theory, of one way of structuring our «cognitive network» as opposed to another, that decided this issue.

It will be noticed that this argument relies heavily upon the phenomenon of scientific revolutions. The increasing acceptance by philosphers of the idea that «revision can strike anywhere», and the aceptance of an idea which is in some ways even more important, the idea that some statements can be overthrown (as Einstein overthrew $p = mv$) *only if* a whole alternative scheme of theory can be produced, and not just by performing experiments, is certainly the result of the great scientific revolutions of the present century, and this goes farthest in the direction of explaining the wide acceptance of «Quinian» views of belief fixation in theoretical science.

Where there is still resistance to accepting the idea that belief fixation is holistic is in the area of «observation language», as Carnap called it, and in what some philosophers call «ordinary language». Carnap himself, while coming to agree with Quine that belief fixation in theoretical science is holistic, seems to have equivocated between conceding that observational language and theoretical language are interdependent, on the one hand, and wanting to maintain that observational language is (1) «completely interpreted» (whereas theoretical language is only «partially interpreted»), and (2) observational language receives its interpretation *prior* to theoretical language, on the other. Michael Dummett, similarly,

seems to concede that theoretical science is holistic, but insists that the sentences of everyday language have canonical verification conditions which are fixed (in a recursive way) by their very meanings. The holism of belief fixation is something which threatens to overturn more than one philosophical applecart unless it can somehow be confined and made harmless.

ORDINARY LANGUAGE AND SCIENTIFIC LANGUAGE

The idea that ordinary language (or «observational language») is unaffected by the instability of scientific theory may be attractive to philosophers of more than one persuasion, but it is just wrong. Anti-holists seem to think that the «operational criteria» used by the layman exhaust the meaning of such a word as *water*. (In the case of *water*, these would include being transparent, colorless, odorless, quenching thirst, not being poisonous, etc.) But it is not hard to see that this is wrong. Suppose there is (as there may be, for all I know of chemistry) a liquid, say «oxyhydroblahblah»», such that it has all the «obvious» properties of water except supporting life: if you drink oxyhydroblahblah it may seem like water (well, maybe there is a slight peculiarity about the taste or feel), but your thirst won't be quenched. A mixture of H_2O and oxyhydroblah-blah will, however, pass as water (no worse than Philadelphia water). Suppose you give a layman a glass of *fifty percent* H_2O and *fifty percent* oxyhydroblahblah. If you tell a layman that it is a glass of 99% H_2O and one percent a harmless something else, he will perhaps still be willing to call it «water». But if you tell him the truth, he won't say it is «water» (although he will say it is *fifty percent* «water»). The fact that what he is given consists of as much as half of a substance which does not occur as a constituent of «normal» water at all certainly debars it from being classed as water. In short, even the layman includes in his concept of water some notion of having a standard or normal «composition»; and this is enough to make his usage of the term *water* dependent on the sorts of facts that it is the business of science to discover.

At one time I thought that the lay and scientific extensions of the term «water» were more simply related than I now think is the case. I thought that the layman's «water» denoted chemically purified H_2O give or take some impurities. But this is *too* simple. A glass of coffee is more than 99% H_2O but the layman won't class it as a «glass of water» —even if he knows that it is more than 99% chemically pure water. He will say that «coffee consists mostly of water», not that it *is* a glass of water. On the other hand, he will probably say that a glass of H_2O with 2% dirt floating in it is a

«glass or muddy water». Our interests determine what sorts of «impurities» debar a quantity of mostly H_2O from being «water» *simpliciter*, and they do this in very complex ways. But this does not effect the point made with the aid of hypotetical oxyhydroblahblah; the scientific extension and the lay extension are different but interdependent. Even the layman's term «water» is not an operationally defined term pure and simple. And this means that the holism of belief fixation will effect even the layman's use.

The layman's concept of water includes having a normal sort of composition, as has been said, but the layman does not himself have the resources or the criteria to determine what the normal composition actually is at an ultimate physical or chemical level. He must defer to experts, past, present or future, or at least assume that there is a fact of the matter as to what the composition is. Of course, the word «normal» itself assumes the existence of a range of different paradigms. If all the water in Lake Michigan turned out to have an «abnormal» composition (relative to other paradigm bodies of water), we woulds perhaps go on considering the liquid in Lake Michigan to be «water» in an «ordinary» sense of the term, as we are said to go on considering whales to be «fish» in an «ordinary» sense of the term. But if a small quantity of putative water of uncertain provenance turns out to have the «wrong» composition (as in the example of a glass of fifty percent H_2O and fifty percent oxyhydroblahblah), we say «that's a mixture of water and something else»; the discovery —which may involve the most recherché scientific theories— that the composition is wrong shows that the stuff isn't (simply) water. «Moderate holism», as Quine calls it, is the right line to take here. The layman's occasional failure to defer to scientific criteria (as when he goes on calling whales «fish»— if, as philosophers claim, he *does* go on calling whales «fish») show that ordinary language and belief fixation are not *as* holistic as scientific language and belief fixation, but it still remains true that scientific revolutions can overturn ordinary language classifications —overturn them to the satisfaction of the *ordinary* speaker, not just the scientist— in particular cases, even if this is a possibility that the ordinary speaker is usually justified in ignoring.

Belief fixation is (in principle) everywhere holistic.

THE MEANING CHANGE ISSUE

Although Carnap came to concede that belief fixation in theoretical science is holistic, he did not give up the idea that *terms* in theoretical science have their own individual «meanings», as Quine urged he should

do. Instead he chose to identify the meaning of a term in a scientific theory with a (complicated) logical function of the theory itself. The «theory», in this conception, was not the mere axiomatic system (viewed as an uninterpreted calculus); rather, it was the axiomatic system together with the given interpretation of the logical words *(and, or, if, not, all, some,* etc.) and the «observational terms». The theory, as Carnap put it, is a «partially interpreted calculus» (the observation terms being —supposedly— «completely interpreted»; the theoretical terms being «interpreted» only by being connected —connected by the theory itself— with observational vocabulary). Various devices from symbolic logic (the «Ramsey sentence» and the «Hilbert epsilon operator») were employed by Carnap in formalizing this idea.

The problem with Carnap's proposal is easily stated: on Carnap's view, every change in a scientific theory involves a change in the meaning of every one of the theoretical terms that the theory contains. Moreover, the «theory» is not just the local or immediate theory in which the term was introduced; because of the holism of belief fixation in theoretical science, it is necessary that the «theory» whose «Ramsey sentence» is taken to fix the meanings of the theoretical terms include all the auxiliary theories which are in practice used in conjunction with the local theory when testable predictions are derived. Thus, in the case of *electron,* for example, Special Relativity, General Relativity, quantum chromodynamics, and the current attemps at «grand unification» all are used in conjunction with «core» electron theory, and all modify the predictions that one derives from the core theory; so all these must be taken as part of the total theory which «implicitly defines» the word *electron,* and each of the stages in the development of this body of theory, including the ones which are currently taking place, *changes the meaning* of *electron,* on Carnp's account.

Consider, for example, the question «Why is H_2O a liquid at 20° C?» Before this question could be answered, quantum mechanics had to be developed. But even quantum mechanics was unable to provide an answer to this question until the discovery of the Pauli Exclusion Principle. This means that, in Carnap's view, «H_2O» changed meaning at least twice before the original question was answered. Or, to speak more precisely, the *original* question was *never* answered —what was answered was a question expressed by words which were homophonic with the words in the original question but not synonymous with them.

This means that a distinction which is crucial in logic and methodology, the distinction between really answering a question and committing a «fallacy of equivocation» has been altered beyond recognition. For we do not regard it as *a fallacy* to answer such a question as this by *developing better theories* —how else should one answer it? Carnap might reply that there are two kinds of «meaning change», a good kind and a bad kind, and

equivocation is a fallacy only when the meaning change is of the bad kind; this would just be to reintroduce the distinction between meaning change in the traditional sense, meaning change properly so called, and meaning change in a «Pickwickian sense».

Not only is it the case that the proposal to take the meaning of the theoretical term or statement to be the theory or a suitable function of the theory involves a radical departure from our «traditional» or preanalytical use of the notion of meaning; it has the effect of robbing that notion of all epistemological interest. If meanings are not invariant under normal processes of belief fixation, then concern with meaning loses its *raison d'etre.*

SOPHISTICATED MENTALISM

So far we have argued that the holistic character of belief fixation cannot be confined to «science». There is another way in which a philosopher might try to block Quine's argument for meaning holism, however, and this way has been vigorously urged in a series of publications by Jerry Fodor. Fodor is fully persuaded that belief fixation —all belief fixation— is holistic, or, as he recently put it, «Quinian». What Fodor challenges is the step *from* holism with respect to belief fixation *to* holism with respect o meaning.

Moreover, Fodor has good reason for thinking that this option is open to him. Quine and Carnap both assume that «meanings» are in some way to be identified with assertibility conditions. If the same sentence occurs in two different theories, what will confirm that sentence will be affected by the difference in the surrounding theories. A sentence in a theory is confirmed indirectly, by confirming the theory as a whole (at least in many cases); if a phenomenon contradicts theory A while being explained by theory B, then a sentence S may decrease in confirmation upon the observation of that phenomenon as long as theory A is still the only theory containing S, but the phenomenon may be regarded later (after theory B has been thought of) as supporting theory B, and, thereby, indirectly increasing the support of the very sentence that it previously disconfirmed. It is for such reasons as this that Carnap counts the change from theory A to theory B as a change in the «meaning» of sentence S and that Quine scraps the notion of «meaning» altogether (in connection with single sentences). But the identification of «meanings» with assertibility conditions (or with rules or theories which determine assertibility conditions) is only binding upon philosophers who wish to retain the idea that *meaning is method of verification.* Quine wishes to retain this idea, and he

is willing to pay the price: to the extent that there are processes which guide us in determining when to accept and when to reject sentences[10], these processes are associated with the whole language, and so, in his view, the «unit of empirical significance» cannot be smaller than *whole language.* Prior to Quine we had been forced to see the sentence and not the word as the primary unit. Since Quine, we have come to see the whole language as the «unit of empirical significance». At least this is how Quine would like future histories of philosophy to read.

Many philosophers, however, including Fodor, do not accept even the «holistic» form of the idea that meaning is method of verification. (I cannot accept it myself). If meaning —even in the case of the whole language or the whole of a speaker's body of belief— is not a function of method of verification or method of confirmation, then it might well seem that there is an easy way to preserve the idea that single words and sentences have determinate meanings, meanings which are invariant under normal processes of belief fixation and theory change, without being led to dispute the holism of the processes of belief fixation and theory change. The general idea which defines this option, the option which I shall call «sophisticated mentalism», is easy to describe: (1) Postulate «meanings» as psychological entities. (But do not postulate that they need to be, at most times, available to consciousness —this is the big difference between contemporary mentalism and the «naive mentalism» of the older empiricists.) (2) Postulate that these entities are, in some way, associated with individual words, morphemes and sentences. (3) Postulate that the step *from* the meaning of a sentence to its assertibility conditions involves the use of general intelligence and background knowledge— which is where the «holism» will come in.

The task of this chapter, and of this book as a whole, is to show that this option has, in a sense, no content. Of course we can say that we «know the meaning» of such a question as «Why is H_2O a liquid at 20° c?» And we say that question means precisely what it seems to mean, that it asks why a certain compound (whose «formula», at the level of high school chemistry, is H_2O) is a liquid at room temperature.

We would not call the discovery of quantum mechanics or the discovery of the Pauli Exclusion Principle, a «change in the meaning» of such a question as this. But something strange happens when we begin to say things like «The question has a meaning», «the meaning has not changed as a result of the discovery of quantum mechanics», etc., in a «theorizing» tone of voice. Taking these «everyday» remarks as *explanatory* puts a burden on the notions of «the meaning» and «change of meaning» that nothing in their ordinary employment equips them to bear. And when we increase the burden by adding such remarks as «meanings are theoretical entities», «meanings are psychologically real», etc., then the notions collapse altogether. For there *are no* isolable states, processes, or objects

—independently and scientifically isolable, that is— that answer to the «meanings» postulated by sophisticated mentalism. At least this is what I shall argue.

THREE CONSTRAINTS ON ANY THEORY OF «CONTENTS»

It is part of our preanalytic notion of «meaning» that different sentences and different words can have the same meaning. If there are psychologically real *entities*, call them «contents», which are associated with individual words and sentences in the way postulated by «sophisticated mentalist» theory and which underly and explain our intuitions about «sameness of meaning» and «difference» of meaning, then synonyms such as the English word «bachelor» and the French word «celibataire» must have the same or closely similar entities of the kind in question associated with them, antonyms («married man»/«bachelor») must have different but structurally related entities (contents) associated with them, and words which are neither synonymous nor in some way related in meaning must have different and structurally unrelated entities (contents) associated with them. Similary, whole sentences which are synonymous must have the same or similar «contents», sentences which are contradictory or otherwise semantically related must have different but structurally related «contents», and unrelated sentences must have unrelated «contents». That this should be so will be our first requirement or constraint on any theory of «contents»:

I. Different contents must, in general, be associated with sentences which we preanalytically suppose to differ in meaning, and the same or closely similar contents must be associated with sentences which we preanalytically suppose to be the same in meaning.

In short, «contents», whatever they may be, must have the correct powers of *disambiguation.*

Ih his celebrated book *Word and Object* Quine attempted to define the notion of meaning, at least for observation sentences, in terms of a behavioristic notion of «prompting assent/dissent» and a neurological notion of «stimulation of nerve endings». The set of all stimulations of the nerve endings that would prompt assent to a sentence S on the part of a given speaker is a behaviorist analogue of the positivist notion of the «sensations that would confirm the sentence»; the set of all stimulations of the nerve endings that would prompt dissent from S is a behaviorist analogue of the «sensations that would disconfirm S»; the ordered pair of these two sets is called the *stimulus meaning* of S by Quine. Quine is clearly attracted to the identification of stimulus meaning with meaning

simpliciter, and, to the present day, he often refers to the stimulus meaning of an observation sentence as its meaning. When he is careful, however, he distinguishes between stimulus meaning and what he calls «analytic meaning». It is analytic meaning and not stimulus meaning that obeys the constraints I am in the process of listing; the stimulus meaning of a sentence cannot and does not determine what a sentence means (in the sense of determining what sentences it is synonymous with or may be paraphrased by), and does not even determine the reference of the sentence-parts. Moreover, it is Quine himself who has argued this. His argument turns on the very constraint I have just listed.

Quine imagines a «jungle language» in which people assent to «*gavagai*» when they see a rabbit and dissent from «*gavagai*» when no rabbit is pressent. More precisely, he imagines that «*gavagai*» has exactly the same stimulus meaning as our English sentence *A rabbit!* (It is important to the argument that *A rabbit* be thought of as a complete —if ungrammatical— sentence, with the force of «Lo, a rabbit!», and not as a noun phrase.) He points out that it does not follow from this supposition that *gavagai* must *mean* «rabbit». Not even if we suppose that *gavagai*, like «a rabbit», can be used either as a noun phrase or as an (elliptical) sentence. For example —assuming that there were a «fact of the matter» as to what gavagai *means*— *gavagai* could mean «undetached rabbit part», or «Rabbithood exemplified again», or «instantaneous time-slice of a rabbit», in full conformity to the facts as specified. If talk of «Rabbithood» became common in English then «A rabbit!» and «Rabbithood exemplified again» could be English observation sentences with exactly the same stimulus meaning. But no rabbit is identical with undetached rabbit parts, or with the universal Rabbithood, or with any instanteneous rabbit-slice. (A «rabbit-slice» is any space-like three-dimensional cross-section of the four dimensional space-time rabbit.)

Rabbit, rabbithood, undetached rabbit part, and *rabbit-slice* are not identical in *denotation*. The number of «Rabbithoods» is exactly one; the number of rabbit-slices is infinite; the number of undetached rabbit parts is probably indeterminate, but certainly much larger than the number of rabbits. Yet as «occasion sentences», «A rabbit!», «Rabbithood», «Undetached rabbit part», and «Rabbit slice» could, I repeat, have exactly the same stimulus meaning. Since sentences which are built up out of parts which differ in denotation are not, preanalytically, identical in meaning, constraint (I) requieres that these four sentences must be assigned different «meanings». Hence meanings (or «contents») cannot be identical with *stimulus meanings*.

Here is a different example to the same effect. I have discovered that many people (mistakenly) believe that albino tigers have no stripes. (Actually, the stripes are faint but still quite visible.) The fact that one can believe «albino tigers have no stripes» and «some tigers have no stripes»

shows that «tiger» and «striped tiger» differ in meaning. Suppose, now, that if there were a stripeless tiger people would hesitate to assent to «Tiger!» perhaps because they weren't sure it wasn't a member of some other species of «big cat». Then the stimulus meaning of «tiger» and «striped tiger» could be exactly the same, even though every speaker knew that a tiger wasn't necessarily a striped tiger!

My next constraint has already figured in the discussion; it is simply that

II. *Contents must remain invariant under normal processess of belief fixation*

To stick to the previous example, let us imagine that all of the members of the speech community learn a great deal about tigers. Eventually, we may suppose, they all become skillful at recognizing tigers without needing to rely on the presence or absence of stripes; if there were a stripeless tiger (or if we simulated one) they would assent to «Tiger»! in its presence without undue hesitation. In this case the «stimulus meaning» of «Tiger!» would have changed, but this is not what we would call a change in the meaning of the word *tiger*. Taking «contents» to be stimulus meanings would violate constraint II (as well as constraint I); for this change in the stimulus meaning would have been brought about by what is certainly a «normal process of belief fixation» —learning more facts about something.

On one reading of his work, Michal Dummett proposes to take the meaning (i.e., the «content») of a sentence to be its canonical method of verification. (He assumes that, in ordinary language as opposed to scientific language, sentences do have stable canonical methods of verification.) However this proposal is also ruled out by constraint II. For the canonical method of verification of a sentence can be effected by discoveries of various kinds (new tests for being gold, or acid, or whatever), or even by technological change (the invention of writing or the telephone). One cannot rule out methods of verification which involve the products of science or technology (consulting written records, in the case of statements about the past; or using a bit of scientific knowledge to determine that one's soil contains too much lead) on the ground that these methods of verification have nothing to do with the *meaning* of the sentences in question, and are *therefore* «non-canonical» without facing the objection that the notions of «meaning» and of being «canonical» have now been explained in terms of each other in such an *immediately circular way* that no progress has been made towards understanding either. Moreover there are *many* statements about the past which cannot be verified *except* by consulting written records; many statements about soils having high lead content which cannot be verified *without* scientific tests, etc. If all of the statements are held to involve «extended uses» of the words they contain, or «non-ordinary» uses, or whatever, then our theory will end up postula-

ting so many «differences in meaning» that it will lose all continuity with our preanalytical intuitions of sameness and difference of meaning. Constraint I will be violated, as well as constraint II.

Finally, since we are investigating the possibility of a *mentalistic* theory —a theory in which the «content» of a word or sentence is a psychological entity which underlies the practice and intuitions of each speaker who is fully competent in the use of the word or sentence, we require:

III. *Their contents must be «associated» with the relevant words and sentences by each speaker who counts as fully competent in the use of the words and sentences*

This constraint rules out theories of meaning on which the contents of theoretical terms which are used by all the members of a speech community *(charged, electron, acid)* are identified with theories (or with logical functions of theories, such as the Ramsey sentence) which are known only to experts. III is, thus, a constraint of publicity (a requirement that «contents», whatever they are, be as widely shared, implicitly, as the competence in the use of the words and expressions whose contents they are) as well as a requirement that «contents» be psychologically real entities. If the content of «charged» is a psychologically real entity present in the head of *every* speaker who is (by normal standards) competent in the use of the word, then that «content» cannot be the sophisticated scientific theory of electricity.

I can now more precisely the claim I made at the end of the last section, the claim that it is the burden of this chapter (and, in a sense, of this book as a whole) to establish: *there are no entities which are plausible candidates for the role of the «contents» associated with our words and other linguistic expressions.* All of the candidates so far suggested by mentalist writers such as Fodor violate a least one of the three constraints; and the suggestion that we just take «contents» as *primtive* entities in psyhology, and *abandon* all hope of identifying them with anything in the neurological or even the computational description of the brain, concedes exactly what I am arguing. For the question is not whether it is useful to talk of words «having meaning» and «not having meaning», «having the same meaning» and «not having the same meaning», etc., in contexts in which it is clear why the question arises and what we are going to do with the answer. The question is whether the idea that meanings are *isolable* events, states, processes, or whatever, with some sort of explanatory role in a genuine scientific theory has the slightest foundation.

AN EXAMPLE

Since the discussion has been on an abstract level, at this point I propose to discuss the «Ruritanian» example from «Computational Psychology and Interpretation Theory». In the example, one of the differences between the dialect of Ruritanian which is spoken in the north and the dialect spoken in the south is that in north Ruritanian *grug* means «silver», while in south Ruritanian this word means «aluminum». We are supposed to imagine that silver is so common in north Ruritania that in the north the pots and pans are made of silver. One might imagine that in Middle Ages «grug» meant «pot metal» in Ruritanian, and that it is the fact that north Ruritanian pots are made of silver and south Ruritanian pots of aluminum that accounts for the meaning shifts that have taken place. In any case, northen children grow up knowing that pots and pans are normally made of «grug» and southern children grow up knowing that pots and pans are normally made of «grug».

In the example, Oscar and Elmer were supposed to be in the same psychological condition (in Fodor's «solipsistic» sense of «psychological condition»), i.e., the same with respect to all internal parameters relevant to language at t_0.

So, although in the adult communities to which they belong, «grug» has two different meanings, it has (in Fodor's sense) the same content for *Oscar* and for *Elmer* at t_0. At t_1 (when they have become adults) the words must differ in content in the two idiolects as much as «silver» and «aluminum» do for speakers of English. Hence the word must *change content* between t_0 and t_1 (for at least one of the children, and presumably for both).

	Elmer	Oscar
t_1	*grug* Elmer = «silver»	*grug* Oscar = «aluminum»
t_0	*same psych. condition*	

However, all that happened between t_0 and t_1 was normal belief fixation (in the model commonly employed in inductive logic, conditionalization of prior probabilities to new information as it continuously pours in). At t_0 both children know that «grug» is a metal, that it is shiny-gray in color, that it tarnishes, that Mother has «grug» pots and pans, etc. By t_1 Oscar knows that grug is called «aluminum» in American English and «aluminium» in British English; that it was briefly very expensive (Napoleon had aluminum jewelry, and it was more costly than platinum), but cecame very cheap; that pots and pans are normally made of grug or of

steel or of copper except in north Ruritania, where silver (called «zilber» in south Ruritanian) is used for pots and pans; that grug comes from bauxite; that «zilber» (which is called grug in the northern dialect) is an expensive metal which does *not* come from bauxite; etc.; and Elmer knows that «grug» is called «silver» in English; that grug is an expensive metal; that grug does not come from bauxite, *etc.* From the «internal» point of view, Oscar would not feel that the acquisition of any of this information was anything but the acquistion of ordinary factual information involving a notion he already had, the acquisition of further facts about «grug». A theory according to which the word «grug» changed its *content* in the course of this perfectly ordinary process of belief fixation flies in the face fundamental properties of the notion of meaning (Constraint II). Moreover, any place we decide to stipulate that the difference in «content» has taken place will be quite arbitrary, and unrelated to actual practices of paraphase and interpretation (will violate constraint I).

In addition, if we postulate that the word «grug» is attached to a formula in «mentalese» (a «semantic representation») at t_0 (in both brains, since the children are in the same psychological condition), then if normal processes of belief fixation do not alter the «semantic representations» of words, the mentalese formula —say, XYZ— will end up having *two different meanings* at t_1. Pushing the problem of meaning back to «mentalese» will have solved nothing, for we will need a theory of meaning of «mentalese» just as much as we need a theory of meaning for any natural language.

My Theory in «The meaning of "Meaning"»

In «The Meaning of "Meaning"» I proposed a *non*-mentalistic theory according to which «meanings» have several components, e.g.

Water

SYNTACTIC MARKERS	SEMANTIC MARKERS	STEREOTYPE	EXTENSION
mass noun, concrete;	natural kind; liquid;	colorless; transparent; tasteless; thirst-quenching; etc.	H_2O (approx.)

In this theory, the meaning of «grug» in Elmer's idiolect is «silver» *from* t_0 *on* and the meaning of «grug» in Oscar's idiolect is «aluminum» *from* t_0 *on*. My theory ascribes the *socially determined extensions* to children's words; if these are different, then we say the words have different «meanings», regardless of what goes on in the children's *heads*. We do not expect a small child to have the knowledge or the skill that we expect an adult speaker to have. We do not expect the average person who uses the word *gold* to have the knowledge or the skill a chemist has. Yet child, average speaker, and chemist all depend on one another and on the things in the environment. If a child is «plugged in» to this network, then we will attribute to its word the meaning and reference the word has in the linguistic community, even if the child is in no position to fix the reference by itself, or to answer questions about the meaning. On this view, the meaning of *grug* is different for Oscar and for Elmer *even at* t_0. In fact, it is correct to say that it means «silver» in Elmer's idiolect and «aluminum» in Oscar's idiolect at t_0.

The theory I put forward in «The Meaning of "Meaning"» did have a mentalistic aspect, however, and this aspect is one I would now give up. In that essay I suggested that we take the psychological component of meaning (the «content») to be the stereotype plus the syntactic and semantic markers. It now seems clear to me that we do not and should not require *sameness* of stereotype before saying that two forms have the same meaning, but rather sufficient similarity, where what counts as «sufficient» is highly context sensitive. In short, I would give up the idea that there is anything which is «the psychological component» of a «meaning». Rather than say, with Fodor, that the «content» of *grug* is the same for Oscar and for Elmer at t_0, I would reject the notion of «content».

More about «stereotypes»

Although I don't think that any philosopher or psychologist would defend the idea that stereotypes can play the role that Fodor assigns to «contents» (Fodor himself has argued that they cannot), I think an examination of some of the reasons that they cannot will bring the entire situation into sharper focus.

Let us begin with a very simple case. No one would deny that among English speakers, stripes are a stereotypical feature of tigers, or that spots are a stereotypical feature of leopards, of that a mane and a «leonine» head are stereotypical features of (male) lions. But there is controversy over the questions, «Is there a *semantical* connection between the notion of a tiger and the notion of having stripes?», «Is there a semantical

connection between the notion of a leopard and having spots?», etc. Why is this the case?

One reason for thinking there is a problem here is certainly a bad reason. Philosophers tended for a long time to write as if «Having "B" is part of the meaning of "A"» means that «All A's have B» is analytic. Now «All tigers have stripes» certainly isn't analytic. So it would follow from this doctrine that the feature of having stripes cannot be part of the meaning of *tiger*, and, by parity of reasoning, that the feature of having spots cannot be part of the meaning of *leopard*, etc. In fact, if Quine is right and there is no analytic/synthetic distinction to be drawn (or if there are no analytic sentences), then it would follow that the notion of «meaning» simply has to be given up. While this is the conclusion Quine himself drew, this seems much too hasty nowadays. What most of us —what, for example, both Fodor and myself— would conclude, is that the notion of meaning has to be separated from the doctrine of analyticity. This way out was, as far as I know, first proposed by Paul Ziff in *Semantic Analysis*. According to Ziff, the feature of having stripes is associated with the word *tiger* whether or not it is in fact true that all tigers, or most tigers, have stripes. Even if it turns out that the stripes are painted on by natives, or a product of a massive deception or illusion, it is still a fact about the English language right now that there is a semantic association between the word *tiger* and the feature of having stripes.

Part of the resistance to this simple (and, I think, very plausible) move on Ziff's part arose from the fact that philosophers had log identified meanings with what they called «concepts». Now «concepts» are things philosophers have theories about —*lots* of theories, some of them going back to the middle ages and even earlier. Having «B is part of the concept A» is very often a way a philosopher says *«All A's have B» is a necessary truth*. Identify «meanings» with concepts and «necessary truth» with «analytic sentence» and *viola!* you get the principle I alluded to above, the principle from which it follows that having stripes is no part of the meaning of *tiger*.

Even if one blocks Quine's move (rejecting the notion of meaning on the ground that that notion presupposes the notion of analyticity) in a different way than Ziff's, by insisting that Quine is just wrong and there *are* some analytic truths, no one would want to claim that «All tigers have stripes» or «All leopards have spots» or «All/male lions have manes» are analytic sentences. So it would still follow that none of the features by which we *stereotypically* distinguish lions, leopards, and tigers from one another is part of the meaning of these words. But then it looks as if *tiger, lion* and *leopard* are going to have exactly the same meaning.

At one time Dretske proposed to drop the idea that natural kind words have meanings. In effect, he proposed to treat these words as traditional theorists often treated proper names, as words with a reference but no

conceptual content. But so much of our language consists of natural kind words, that to drop the notion of meaning in connection with these can only be the first step to Quine's position, the position that the notion of meaning must be abandoned altogether. Moreover, similar difficulties arise with other classes of words. Is it *analytic* that pencils are artifacts? (Perhaps they are really organism.)

Any notion of meaning which has continuity with our preanalytic notion must be able to «disambiguate» *lion, tiger,* and *leopard.* And the only plausible way to do this is the way we «preanalytically» do it, the way we «intuitively» do it: to say that *tiger* has a different meaning from *leopard* because, *inter alia,* the feature of having stripes is associated with *tiger* while the feature of having spots (rather than stripes) is associated with *leopard.* So far we have to follow Ziff. Or so it seems to me.

The mistake I made in *The Meaning of «Meaning»* was to try for too much definiteness in the notion of a «stereotype». Even if we agree with Ziff that tigers are stereotypically striped and leopards stereotypically spotted, it does not follow that there is any such thing as «the» stereotype of a tiger, leopard, or, for that matter, of a house cat; it does not follow that there is some invariant object (the same for every speaker who is competent in the use of the word, as required by our constraint III) present in the mind/brain of speakers who know the word, or any word with «the same meaning».

Here is an example in connection with the word *cat,* or rather, the Thai word dictionaries translates as *cat,* the word (believe it or not!) *meew* (pronounced as a Frenchman would pronounce *meu):* I have learned that all the cats one sees in Thailand (formerly Siam) are Siamese cats. A person growing up in Thailand has quite a different stereotype of a cat from the one we have. In fact his stereotype of a «meew» is our stereotype of a «Siamese cat».

How, then, should we translate *meew?* If we translate it as «Siamese cat» we convey the correct stereotype but the wrong extension; for the Siamese also refer to European cats and American cats as «meew». So it would just be plain wrong to say that «*meew* means "Siamese cat"». (Just as it would be plain wrong to say that *tiger* and *striped tiger* are synonyms on the ground that the stereotype is the same.) Stereotype may be a «component» of meaning, to use the terminology I employed in «The Meaning of "Meaning"», but there are other components as well. Should we translate *meew* as «cat»? Well, that is what we actually do. We translate in such a way that the extension comes out right, even though the stereotype is not preserved. Yet there are cases in which we say that terms with the same extension do *not* have the same meaning. If a language had terms, say, *cardiolate* and *lungolate,* for «creature with a heart» and «creature with a lung», we would not say that *these* were

synonyms even though they might well have exactly the same extension. In short, there does not seem to be any general rule.

A further problem (one which was mentioned in «The Meaning of "Meaning"») is that stereotypes are not just images (or «perceptual prototypes», in psychological jargon) but beliefs stated in words (a hydrogen atom contains an electron and a proton», «aluminum is a metal»). To know what someone's stereotype of something is requires being able to interpret that person's language: a theory which takes the notion of a «stereotype» as primitive cannot *explain* interpretation.

Suppose we attempt to evade this difficult by restricting the notion of a stereotype to features than can be *pictured.* (And suppose we imagine that psychologists have discovered a reasonable criterion for determining whether a feature of something is «stereotypical» form a given subject, in this sense.) Even so, to take every difference in «stereotype» (in such a sense) to be a difference in meaning would depart totally from actual practice. Not only would it turn out that *meew* does not have the same meaning as *cat,* but worse, if Tom next door does not include the whiskers of a cat or the milk-drinking as part of *his* stereotype and Jane does, then the meaning of the word is different for Tom and Jane. Of course, many «folk philosophers» *say* just this: «no word has quite the same meaning for two different people»; but that's *not* how we actually *use* the notion of «meaning» when we *aren't* being folk philosphers. In fact, all the sophisticated mentalists I know *attack* the idea that meanings simply *are* stereotypes (and rightly so; can we even make sense of the notion of a «stereotype» in the case of such words as «mind», «esprit», «Geist»?); but if someone were so ill-advised as to hold this view, we should just advise him to *drop* the word «meaning» and to talk about the role of stereotypes in communication. (Of course, they aren't invariant under belief fixation, even barring scientific revolutions, they don't determine paraphrase relations...)

Thus, the two extreme courses —namely, to deny that stereotypes play *any* role in fixing the meanings of words, and to count *every* difference in stereotype as a difference in the meanings of words— both lead to theories of meaning which disastrously violate our three constraints. The course we actually follow in interpretation is a middle one between these two extremes. First of all, when short words have exactly the same extension, we tend to treat them as synonyms regardless of what's «in the head». This is why mentalistic theories of an M.I.T. type are in trouble from the start: they try to *factor out* extension; but extension is what most strongly guides us in interpretation, especially when there is no specific context to guide us. (Yes, I know about «creature with a heart» and «creature with a lung» having a different meaning but the same extension; but the *parts* of these phrases don't even have the same *extensions.)*

In the second place, when we consider factors beyond the extension, we *do* consider stereotypes (those stripes *are* somehow connected with the

meaning of «tiger»), but what we are concerned with is not *identity* of stereotype (however that might be defined) but *sufficient similarity*. And there is no general rule for deciding when two stereotypes are sufficiently similar; it depends on the particular context, including the reasons why someone wants to know what a word means and what he is going to do with the answer.

To sum up: I have not argued that meaning is indeterminate (although Quine would so argue). What I have argued is that *when* meaning is determinate it is *no one thing* that makes it so. The «standard meaning» of *meew* in Thai is *cat;* not because the *same* entity is in the heads of Thai and English speakers, but notwithstanding the fact that there are psychological differences. To say that meanings are the «same» is just to say that it is reasonable, in a given context, to ignore differences in belief, attitude, stereotype, on the part of the speakers or writers in question. The «sameness» of the meanings is not the «sameness» of two objects (two protons or even two neckties), but just the reasonableness of ignoring the differences, including the differences in the «psychological processes».

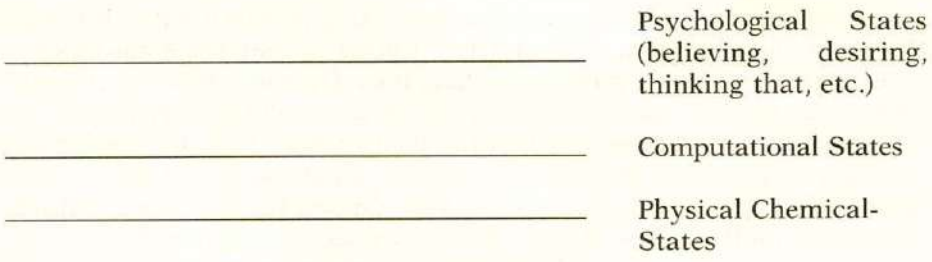
MY PRESENT PICTURE

I shall close by describing the bearing all this has on the question of «functionalism». The functionalist picture was a two-level picture,

F_1 F_2 F_3 F_4 F_5 F_6	Computational (software) States
	= (Psychological States)
S_1 S_2 S_3 S_4 S_5	Physical-Chemical States

in which each psychological state was identified with a computational state which could be realized by a large (potentially, an infinite) number of physical-chemical states. Later is was realized that not only could a given computational state be realized in infinitely many disparate ways whose relationship to one another could not be seen at the physical-chemical level, but that one physical state could also be the realization of an infinite number of different computational states; that a physical-chemical state does not have an «intrinsic» computational description. Thus the relationship of the states on the two different levels is many-many.

This was an *anti-reductionist* picture insofar as it denied any direct reducibility of psychological properties to physical-chemical properties. But there was still supposed to be a reducibility of psychological properties to computational ones. Today, in view of the sorts of considerations just reviewed, I would move to a three-level picture:



with many-many relations between any two levels. In all this, what we are gradually seeing is a breakup of the reductionist picture.