Philosophical Perspectives, 11, Mind, Causation, and World, 1997

PUTNAM'S PARADOX: METAPHYSICAL REALISM REVAMPED AND EVADED*

Bas C. van Fraassen
Princeton University

Hilary Putnam's argument against metaphysical realism (commonly referred to as the "model theoretic argument") has now enjoyed two decades of discussion.[1] The text is rich and contains variously construable arguments against variously construed philosophical positions. David Lewis isolated one argument and called it "Putnam's Paradox".[2] That argument is clear and concise; so is the paradoxical conclusion it purports to demonstrate; and so is Lewis' paradox-avoiding solution. His solution involves a position I call "anti-nominalism": not only are classes real, but they are divided into arbitrary and 'natural' classes. The natural classes 'carve nature at the joints', being (as other philosophers might say) the extensions of 'real' properties, universals, or Forms.[3] Thus the argument was turned, in effect, into support for a metaphysical realism stronger than Putnam envisaged.

I will offer a different way to look at Putnam's model theoretic argument. If we insist on discussing language solely in terms of a relation between words and things, we may well be forced into a metaphysical realist point of view, on pain of paradox. But on the level of pragmatics, in a discussion of language that also addresses the roles of user and use, the air of paradox dissolves all by itself. I shall also try to show how we can resist Lewis' argument, derived from Putnam's, for anti-nominalism. No metaphysical postulates will be needed to avoid the threatened disasters of self-understanding.

1. Putnam's Model-Theoretic Argument

Putnam's argument is very simple, and it is clearly right. What is not clear is just what it is right about. Here is the argument as presented to the APA in 1976:

> So let T_1 be an ideal theory, by our lights. Lifting restrictions to our actual all-too-finite powers, we can imagine T_1 to have every property *except objective truth*—which is left open—that we like. E.g. T_1 can be imagined complete, consistent, to

> predict correctly all observation sentences (as far as we can tell), to meet whatever 'operational constraints' there are... , to be 'beautiful', 'simple', 'plausible', etc... .
>
> I imagine that THE WORLD has (or can be broken into) infinitely many pieces. I also assume T_1 *says* there are infinitely many things (so in *this* respect T_1 *is* 'objectively right' about THE WORLD). Now T_1 is *consistent*...and has (only) infinite models. So by the completeness theorem... , T_1 has a model of every infinite cardinality. Pick a model M of the same cardinality as THE WORLD. Map the individuals of M one-to-one into the pieces of THE WORLD, and use the mapping to define relations of M directly in THE WORLD. The result is a satisfaction relation SAT—a 'correspondence' between the terms of [the language] L and sets of pieces of THE WORLD —such that theory T_1 comes out *true*—true of THE WORLD—provided we just interpret 'true' as TRUE(SAT). So whatever becomes of the claim that even the *ideal* theory T_1 might *really* be false? (Putnam 1978, pp. 125-126)

Let me begin with a quick construal of this argument, similar to that provided by Lewis (which I shall discuss at some length below) and Elgin.[4]

Questions of truth and falsity cannot arise for a theory unless the language in which it is formulated is more than mere syntax: the names and predicates must have specific extensions, and so forth. A function which assigns those extensions to the non-logical words is an *interpretation*. So in the first instance, the question of truth makes no sense unless we fix an interpretation for the language. If we only have criteria to narrow down the interpretation, then the theory is true simpliciter precisely if it is true relative to some interpretation that meets those criteria.

Now here is the punch: if the criteria we lay down are solely 'internal'—to the effect that certain sentences must come out true (since they reflect our intentions about how to use the language), then practically every theory is true.[5] After all, the extensions can be assigned in some way so as to do the job. Perhaps we can accept this consequence, happy to have found truth so easily.[6] If not, it seems we had better postulate 'external' constraints, binding or gluing our words to things quite independently of our intentions and desiderata, and giving factual content to our theorizing.

Being extraordinarily general in form, Putnam's argument seems almost entirely independent of the character of the language in which the theory in question is formulated. One might accordingly surmise that the conclusion is easily refuted by counterexample. However, all that counterexamples can do is to shift the burden to the content of those theoretical and 'operational' constraints which Putnam admits.

Let me illustrate this with as simple an example as possible. Suppose language L has only one non-logical term, namely "water". Very little can be said in L: we can say that there is water, how many things are water, how many are not. As axioms for theory T_1 take the set of sentences which say in effect, for each natural number N, that there are at least N things that are water, and at least N things which are not water. This theory is complete (within L) and has only infi-

nite models. But if, for example, every sample of water is a finite collection of water molecules and if there are only finitely many such molecules in all, then T_1 is false on any interpretation of L on which "water" denotes water.

Putnam's argument can be saved from refutation by such counterexamples, but only in one way: by insisting that the constraints admitted do not suffice to fix the reference of "water". To give an example, suppose we treat "water" as an alien term. Let us now require T_1 only to imply all actual observation sentences which classify things as water or as not water. In that case, if only finitely many such sentences are uttered, then there will be an infinite class of things left (in an infinite model or world) which an interpretation may place inside or outside the extension of "water". On these suppositions, the extension of "water" is not sufficiently fixed by the operative constraints to rule out interpretations which satisfy T even though in reality only finitely many things are water. So read, we can easily understand Lewis' complaint that "the reason given [in Putnam's argument] is, roughly, that there is no semantic glue to stick our words onto their referents, and so reference is very much up for grabs" (1984, p. 221). If Putnam's argument is a *reductio ad absurdum*, it must have premises ruling out a view of language (surely common enough, and not very audacious?) according to which the extension of "water" is fixed, settled, and specific.

If we read the conclusion only a little differently, however, we'll see nothing troublesome at all. Putnam's argument applies quite obviously to languages subject to very few constraints on reference. To give this bare bit of logic some application we must imagine ourselves in the role (made so familiar by Quine) of alien and alienated, possibly extraterrestrial, anthropologists. They study the recordings of a language found only in, say, a small tribe on the upper Amazon or Congo. In that case, what does Putnam's argument reduce to absurdity? Answer: that any such disadvantaged anthropologist can arrive at a unique translation of that language, with the extensions of its words fixed, settled, and definite for them, even relative to all such evidence obtainable in the long run. But that is too obvious; we did not need a highpowered argument to convince us of *that*! Where is the punch?

To feel the troubling force of Putnam's argument, we must convince ourselves that *really* we are in the same position with respect to our own language as those anthropologists are with respect to their object of study. Begin by saying that Putnam established his conclusion for a very large class of languages, perhaps even all languages (if that makes sense). Add that our own language—the very language in which we state his argument and develop our scientific theories—is one of those. Infer that his conclusion is true of our own language. There you have it!

Trying to block this universal instantiation looks absurd, doesn't it? But the conclusion, that our ideal theories cannot be false, looks equally absurd. So what can we say in retort? To our dismay it seems we have no recourse but to say: Putnam's argument pertains only to languages lacking the semantic glue to stick

their words firmly to their referents—so, our language must be different. We seem to have no other way out.

But that is not so.

2. Paradox Lost

I displayed the original text, for two reasons. The first is that it is couched in terms of the so-called syntactic, axiomatic view of scientific theories. A theory is identified as a set of sentences in a vocabulary which is divided into 'observation terms' and 'theoretical terms'. This point may be merely incidental to Putnam's argument, but will be important below when we examine Lewis' response in detail.

The second and more important reason is to highlight its anthropomorphic way of talking about mathematical entities. Putnam commands us: "Pick a model M... . Map the individuals of M one-to-one into the pieces of THE WORLD". When stating what I took to be a common or obvious construal of the argument, I took care to keep some of this 'human' language about the interpretations. Note well that the argument, and the metalogical theorems to which it appeals, are solely about the existence of functions. Nevertheless, they are couched in the discourse of physical manipulation. That way of speaking, as if we humans are actually carrying out specific tasks, may be harmless when we are doing pure classical mathematics. Is it really so harmless here?

What Putnam (apparently) commands us to do here may or may not be possible. He has said that model M and THE WORLD have the same cardinality. Therefore there certainly exists a one-to-one map between them—indeed, very many such maps exist. But can we identify or pick one of them?

Here is an analogous example. Consider a geometric object, a sphere in Euclidean space. Can we coordinatize the surface of this sphere? Our first inclination is to say Yes. There certainly exist many functions that map the surface points into triples of real numbers in the right way. But I asked: can WE do this? The answer is No, for this object has perfect symmetry. We would like to give one point the coordinates (0,1,0) and call it the North Pole. But this point is not distinguished from any other on the sphere, so we cannot do it. Of course, if the sphere is not a mathematical object but, say, the Earth, then we can do it. The reason is that we can independently identify a spot to designate as North Pole. Similarly, we may be able to do it if the sphere is already related to some other mathematical object, has some functions defined it, and so on—for the same reason. But we cannot do it unless we have some independent way to describe points on that sphere.[7]

Does this distinction—between the existence of a function and our being able to carry out the mapping—matter here? It does indeed if we want to apply Putnam's argument to our own theories formulated in our own language. We shall be able to grasp such a theory if we can grasp even one interpretation of the language in which it is formulated (i.e. our language). (This view of understand-

ing language seems to be implicit in Putnam's argument; we need at least to count it among the premises of the reductio.) Well, we can grasp an interpretation—i.e. function linking words to parts of THE WORLD—only if we can identify and describe that function. But we cannot do *that* unless we can independently describe THE WORLD.

So to Putnam's model theoretic argument, if meant to be applied to our own language, we must respond with a dilemma:

> (A) if we cannot describe the elements of THE WORLD, neither can we describe/define/identify any function that assigns extensions to our predicates in THE WORLD;
> (B) if we can describe those elements then we can also distinguish between right and wrong assignments of extensions to our predicates in THE WORLD.

The extent to which we can describe or identify specific elements of the world is not at issue here. If we can describe the world to the minimal extent of being able to *use* the word "water" or "cat", we can say that an interpretation, if applied to our language, is wrong unless it assigns water as referent to "water" and the set of cats to the words "cat". To specify that the set of cats is also the extension of "gatto" and "kat", or the set of groundhogs the extension of "woodchuck", needs more. Each of us has his limitations there; but that is not what this argument is about.

This paradox-dissolving response does not transpose to those anthropologists studying recordings of an alien language. We cannot lessen their methodological burden in this way. It works only when the language *we* are discussing is *our own* language, the language on which we implicitly rely throughout. That is the only place where discussion at the pragmatic level does not effectively reduce to an equivalent semantic account. Accordingly, what I have just done is not to refute Putnam's argument. I have only dissolved the paradox that results from an uncritical application—whether or not sanctioned by what Putnam calls "metaphysical realism"—to our own language.[8]

If we deal with our own language as if it were an alien syntax, we will certainly land ourselves in paradox. For then, if we try to discuss truth, we can get no further than truth-relative-to some interpretation. But for texts in our own language the two cannot be equated. Putnam simply is not at liberty to say so blithely "provided we just interpret 'true' as TRUE(SAT)", if what we are concerned with is an attribution of truth to a theory stated in our own language. Humpty-Dumpty alone has ever thought differently.

Only if we lose sight of this pragmatic aspect of Putnam's argument shall we be led into the fallacy of seeing a gap to be filled by metaphysics. For then the adequacy of a theory must seem to derive in part from the adequacy of the language in which it is stated. This adequacy must then in turn seemingly derive from something that makes our language especially blessed among languages:

privileged in some objective, use-independent way. The road into this fallacy is clear. It should be equally clear how we can refuse to enter upon it.

3. What If We De-Anthropomorphize?

In my discussion of Putnam's argument I drew attention to a certain anthropomorphic way of speaking about mathematics—a picturesque, analogical form of discourse. But isn't that merely incidental to Putnam's argument? If the same argument is stated in more 'literal' language, does it not go through just as well, and doesn't my diagnosis and cure lose their very basis?

Well, let us see. Let T be a theory that contains all the sentences *we* insist are true, and that has all other qualities we desire in an ideal theory. Suppose moreover that there are infinitely many things, and that T says so. Then there exist functions (interpretations) which assign to each term in T's vocabulary an extension, and which satisfy T. Does it not follow that T is true, that is, that T is a true description of what there is?

Now I have avoided the 'action' discourse that depicts us as selecting, identifying, or constructing a specific interpretation, and the argument seems to have as much punch as before. But we very clearly do not have to say Yes to the concluding question—and that for the same reason as before. For we may point out that all those functions may be defective in some way not ruled out by the given. Any such function may, despite satisfying T, fail to give the set of green things as extension to "green", or the set of cats to "cat". It may also fail to satisfy "The total number of dinosaurs was a multiple of 17" if and only if the total number of dinosaurs was a multiple of 17. If in fact all those functions that satisfy T have some such defect, then T is not true.

This objection does not work if T was not a theory stated in our language. But if it is not, then we do not have a case that need raise any philosophical eyebrows. In stating my objection, it is true that I relied on our understanding of our language; but so I did while stating the argument, after all.

Could we take Putnam's argument in an Anselmian, "even the fool sayeth in his heart", way? That would be to take it somewhat as follows:

> Let us rely on our understanding of our own language in order to state the argument. Having reached its conclusion, we note that it throws grave doubt on the reliability of that language. Thus we have performed a sort of reductio ad absurdum of the belief that we can rely on our own understanding of our language in use.

Well, we can take it so, but then it is very unsuccessful. For if we are allowed to rely on and use our own language to this extent, we can point out that the argument is not valid. The conclusion that T_1—a theory stated in our language in use—is true *does not follow* from the given, which entails only that T_1 is true

under some interpretation. If all the interpretations which satisfy T are defective in not assigning all and only cats to the extension of "cat" (mutatis mutandis for any other predicate) then T is false. So the argument does not succeed in throwing grave doubt on the reliability of our language in this convoluted Anselmian way either.

The reader may reasonably feel that my putative dissolution of Putnam's paradox needs further support. Specifically, I've brought into the fray certain views of language here presented only very briefly and sketchily. Opposed are views so well entrenched (at least in 1976) that Putnam could appeal to them implicitly and expect his audience to go along. In the concluding section of this paper I shall try to provide some of the needed further support by discussing pragmatic tautologies, interpretation, and translation. But at this point, for the next four sections, I will examine an entirely different (realist) diagnosis of Putnam's argument, which entailed a very different conclusion.

4. Lewis' Diagnosis of the Argument

How could Putnam's argument have been read as, first of all an objection to any sort of Realism, and secondly as providing the clue and support for an espousal of a specific Realist (anti-nominalist) philosophical position?

To find out, we must examine Lewis' analysis of the argument. I'll outline what I take to be his strategy, examine how that strategy is implemented, and then show where we can take issue with it. To begin, Lewis exhibits Putnam's model-theoretic argument as putative demonstration of "Putnam's incredible thesis", to wit, the denial that any "empirically ideal theory, as verified as can be, might nevertheless be false" (p. 221). Next he shows that Putnam's demonstration suffices also to prove that Global Descriptivism (I'll explain this below) implies Putnam's incredible thesis. Thus we are faced with a trilemma:

> So global descriptivism is false; or Putnam's incredible thesis is true; or there is something wrong with the presuppositions of our whole line of thought. Unlike Putnam, I resolutely eliminate the second and third alternatives. The one that remains must therefore be the truth. Global descriptivism stands refuted. (p. 224)

What has intervened in the meanwhile is an argument to show that a certain inescapable question about language must be answered by asserting either Global Descriptivism or some rival thereto. Any such rival will entail anti-nominalism; and so this trilemma, with its one horn remaining unrejected, leads to:

> There must be some additional constraints on reference: some constraint that might, if we are unlucky in our theorizing, eliminate *all* the allegedly intended interpretations that make the theory come true. (p. 224)

So far Lewis' strategy. Let us now examine how discussion of Putnam's argument brought us into this fateful confrontation with Global Descriptivism.

What exactly is Global Descriptivism, and to what question is Global Descriptivism an answer? Lewis begins by outlining an idea in philosophy of language which he calls "local descriptivism". It answers the question how newly introduced terms are to be understood in a growing language. The answer has in effect two parts. Under unfelicitous circumstances they remain defective in certain ways. If the conditions are felicitous, they refer to something identified by a definite description in the 'old' language, the language as it was before the new term was introduced. With seven listed amendments to the original simplistic idea, this answer is accepted by Lewis. However, it does not answer a more ambitious question:

> a local descriptivism is disappointingly modest. It tells us how to get more reference if we have some already. But where did the old language get *its* reference? (p. 223/4)

Here we have the genuinely crucial, central question at the heart of the entire discussion. How did our language as a whole—how did all our terms—acquire reference?

According to Lewis, acceptance of local descriptivism leaves this as inescapable open question and its answer must either be Global Descriptivism ('more of the same') or some rival thereto. Global Descriptivism is formulated as follows:

> The intended interpretation will be one, if such there be, that makes the term-introducing theory come true. (Or: the intended interpretations will be the ones, if such there be, ...with indeterminacy if there are more than one.) But this time, the term-introducing theory is total theory! Call this account of reference: *global* descriptivism. (p.224)

How does this connect with local descriptivism? Well, the latter could be paraphrased as: the new terms appear in new theories, and they are to be understood as referring to things which are as described in the old terms of that new theory, so as to make that theory come out true—if possible. So Global Descriptivism is something like: all language is to be understood as referring to things in such a way that the total theory (the totality of our beliefs or assertions) comes out true.

But now Putnam's argument can be used to show that such a way of understanding our language will be available, pretty well regardless of what that total theory is. In other words, Global Descriptivism implies Putnam's incredible thesis. Thus we have arrived at the trilemma stated above: either Global Descriptivism is false, or Putnam's incredible thesis is true, or there is something wrong with our whole line of thought. Looking back to our last citation, we see that to deny Global Descriptivism is to assert that the intended interpretation, what our language really means, must be constrained by something else than our inten-

tions, verifications, and other desiderata that we ourselves impose—by something other than us, by the world. That is anti-nominalism, however you cut it.

Now I have outlined Lewis' strategy, and I have explained its implementation by showing how he arrives at the trilemma and how he goes from that trilemma to anti-nominalism. It remains now to show that the appearances here created were very deceptive.

What is actually the case is this. Lewis arrives at the trilemma via the argument that local descriptivism is successful in its own domain, but leaves an open question which must be answered by either Global Descriptivism or one of the rivals thereto. In actual fact, if local descriptivism is successful, it leaves no such open question. Secondly, the version of local descriptivism which Lewis had advanced earlier, in his theory of science, already required anti-nominalism for its tenability and success. Thirdly, that version of local descriptivism is only disputably successful, and not compelling outside the context set by certain philosophical assumptions.

The first of these contentions I shall support in the remainder of this section. The second I take only to recount an acknowledged part of Lewis' position, not an objection. I'll make it plain in the next two sections. Then I will give my reasons for the third, that is, for holding that we need not accept that part of his position. The general questions about language to which local and global descriptivism give putative answers I'll return to in the final part, and I'll argue that indeed, "there [was] something wrong with the presuppositions of [this] whole line of thought".

So let us take a look at this question that local descriptivism supposedly leaves dangling. Assume that local descriptivism is successful in its own domain, that is, that it explains the meaning of new terms that enter our language—e.g. through scientific theory innovation—by relying on descriptions formulable in the old vocabulary. We must think of this as an account applicable to every stage in which language undergoes such innovation. Lewis reflects on this supposition, as we recall: "a local descriptivism is disappointingly modest. It tells us how to get more reference if we have some already. But where did the old language get *its* reference?" (p. 223/4)

The answer is not to be of the "turtles all the way down" type. Shades of Aquinas! Why not? If local descriptivism gives a good answer for the terms introduced in 1900 AD, why not also for those introduced in 1900 BC or 19,000,000 BC? Suppose there was a stage when our ancestors introduced new terms, and these terms were successfully used to refer to things. If those were not things that could have been described sufficiently well in the 'old' language to give local descriptivism its purchase, then it follows that local descriptivism is not tenable. So, tentative conclusion: we cannot say *both* that local descriptivism is a tenable answer to its own question, *and* yet that it leaves the global question unanswered.

This is not a peculiar point about local descriptivism. The same point applies to any theory, empirical or philosophical, concerning how a language grows,

acquires new resources and new vocabulary, and how its newer stages are as successfully adapted to its functions as the old was. If such a theory is to be adequate at all, it must account for the very early growth of linguistic behaviour among our earliest language using ancestors. It must also account for apparent radical changes that make hermeneutics of ancient and medieval texts so challenging. But if it is adequate, it leaves nothing unanswered; the language as a whole grew successfully because each stage grew properly, or well enough, out of its antecedents.

How does this affect Lewis' argument? Perhaps not at all, except to highlight some assumptions behind the trilemma. Lewis does not rely on the success of local descriptivism, but only on the failure of global descriptivism. The question "where did the old language get its reference?" (p.224) is understood not as a historical question about a particular stage (the language of 1900, say) but as being about all language—and the form of answer is then glossed as "the intended interpretation(s) will be the one(s) such that...". Are we to give this answer in the same language in which it is posed? If not, in what language?

I wonder if we are to think that our answer should be the same for the actual history of humankind as it would be for a science-fiction case like the following:

> on a certain planet, newly discovered by our space explorers, spontaneous generation occurred of an intelligent language using species. Within a period of days inanimate crystals transformed naturally into living, moving, speaking beings. At first hearing, it certainly seemed as if they were already equipped with a large array of beliefs about the world, a total theory of their own. The question was: how to interpret what they were saying? Should we adopt the theory that what they were referring to and were saying about it must be such as to make their total theory true?

A wonderful problem! It should definitely be posed to the special creationists teaching in the Bible Belt. I'm not sure how this problem could arise, for how could it seem to us that they were speaking and had a theory? Well, perhaps it sounded exactly as if they were speaking German and propounding Lysenko's biology. That would give us a salient initial interpreting hypothesis—not necessarily the one to maintain, of course. In view of the paranormal circumstances, it is clear that local descriptivism will not work here. Nor will a more empirical theory that assimilates language use to tool use and looks for selective pressures, fitness, adaption to environment, and so forth. There were no preceding behaviour, dangers coped with, or prior needs to be satisfied.

The little dialectic about how local descriptivism is not an answer to certain questions—to which global descriptivism is an unacceptable answer—is to me very suspicious. It's point, subliminally made, I suspect, is that after questions are taken as answerable empirical questions, there remain questions of the same apparent form which go beyond all empirical inquiry and need philosophical answers. Local descriptivism is a rather fanciful stand-in for an empirical theory of

how language grows. Actually it may not be adequate, namely if it cannot account for growth in which the linguistic resources become genuinely richer, and if such growth has actually happened as the human race evolved (nor of course if something like our science fiction story happened instead). Be that as it may, an adequate theory of that ilk would leave nothing unanswered about the growth of the whole language. It would need neither global descriptivism nor some rival thereto as its supplement. The refutation of global descriptivism refutes something for which there was no call in the first place.

It appears therefore that re-examination of Putnam's argument never does bring us into that fateful dilemma. We could stop our examination of the realist reaction at this point. But realism is hardy, and we had better inquire into the second and third issue I announced above.

5. "How to Define Theoretical Terms"

The transit from Putnam's argument to the trilemma began with a short position statement to the effect that local descriptivism is successful in its own domain. Only the failure of global descriptivism was taken as basis for the further move to anti-nominalism. But in actuality, the local descriptivism which Lewis had advanced earlier will turn out—in the light of Putnam's Paradox—to require reliance on anti-nominalism. Thus rejection of the question that evokes global descriptivism may not get us very far. If local descriptivism is to be accepted then it will bring anti-nominalism along with it. To avoid being outflanked in this way, we must therefore examine and find reasons to resist local descriptivism as well.

The story begins with David Lewis' paper "How to Define Theoretical Terms". This addresses a problem inherited from that stage of logical positivism which divided the vocabulary of science into two parts: observational and theoretical. No such division can be tenably maintained, but Lewis points out that there is a problem nevertheless. When a new theory is introduced, with new terms which cannot be explicitly defined by means of the retained old vocabulary of science, the question is still: how shall we understand the new theoretical terms? Lewis proposes accordingly to focus on a hypothetical particular time t, at which a new theory is introduced. Relative to this time t we can make a historical division of the vocabulary into Old vocabulary and New terms.

To forestall misunderstanding please note that the meaning of the Old vocabulary is fixed. The Old terms are assumed to be completely understood at the time in question:

> by 'understand' I mean 'understand'—not 'know how to analyze'
> Let us assume that the [Old terms] have conventionally established standard interpretations[9]

Can we view the newly introduced theory as formulable entirely in the Old terms, in principle, so that it will be intelligible to the community of Old vocabulary speakers?

As first step we consider the proposal to replace a theory by its Ramsey sentence. That sentence is formed by replacing all the New terms by variables of appropriate type, and then existentially quantifying those variables. Without loss of generality (let us assume) the New terms will all be predicates, so this is higher order quantification. As example consider a very simple theory with three new theoretical terms:

> Water consists of hydrogen atoms and oxygen atoms.

The Ramsey sentence of this theory is:

> There exist three properties such that water is composed of entities which have the first and third property and entities which have the second and third property.

A caricature of an example, of course. But the crucial point is the same as for any more extensive theory: all consequences of the original theory which can be stated entirely in the Old vocabulary are also consequences of its Ramsey sentence.

Ramsey's idea can be put roughly like this: when scientists introduce a New term like "hydrogen" they are simply following the mathematicians' habitual play with anaphora. "If a curve passes through points (0,0) and (2,2) then it must share some point with the line $x = 1$; let us call that point p." Which point? There are many points where the curve may cross the line! There can't be a pretense of having denoted a specific point. But the mathematician doesn't care, for he is really reasoning 'within the existential quantifier', and the name "p" won't occur in his conclusion. In just the same way, one might say, the scientists' real conclusions do not involve those New theoretical terms—the consequences stated in Old terms alone are what constitute his predictions and whose truth is the bottom line for his theory.

This idea is attractive only if what is predicted and tested, at this time, is just what is stated in Old terms alone. That presumption carried a large part of the appeal of the logical positivist theory of science which I mentioned above. We'll let it stand for now, and return to it below. There is a more immediate problem. Agreed that replacing a theory by its Ramsey sentence eliminates the New terms—but how is that a step toward understanding those terms?

Logically speaking, the Ramsey sentence is a great deal weaker than the original theory and the pattern of properties and relations which it describes can normally be realized in many different ways. There may be many properties and relations which together will play the roles that the theoretically introduced properties and relations are meant to play.

Lewis' proposal is to replace the theory not with its Ramsey sentence, but with (in effect) a combination of two postulates: the Ramsey sentence plus a uniqueness postulate. The uniqueness postulate says that there is a unique selec-

tion of properties and relations which realize the pattern described in the Ramsey sentence. Relative to the combination of the two, the New terms will all be explicitly definable. Via those explicit definitions, the replacement will logically imply the original theory. Thus this replacement, by (in effect) the Ramsey sentence plus uniqueness postulate, replaces the theory with something logically stronger. The new idea is therefore that we should construe the scientist as really asserting somewhat more rather than less than what he seems to be saying explicitly.

It is not easy to illustrate this. Our simple 'water theory' example would be recast, in effect, as:

> There exists three and only three properties such that water is composed of entities which have the first and third property and entities which have the second and third property.

This sentence seems obviously false, while the original theory, many people would say, is true. But of course the proposal can be defended. The blame may be put on my choosing too small a part of science as my sample theory. If I could have chosen the sum of everything current physics has to say about water, the Ramsey sentence would have been very much more complicated. No one could have said at first glance that it has multiple realizations (if any at all).

Here is Lewis' succinct argument for understanding scientific theories as implicitly (or provisionally?) involving a uniqueness postulate of this sort:

> Is there any reason to think that we must settle for multiply realized theories? I know of nothing in the way scientists propose theories which suggests that they do not hope for unique realization. And they know of no good reason why they should not hope for unique realization. Therefore, I contend that we ought to say that the theoretical terms of multiply realized theories are denotationless.
>
> Many philosophers do seem to think that unique realization is an extravagant hope, unlikely in scientific practice or even impossible in principle. Partly this is professional skepticism; partly it is skepticism derived from confusion that I shall try to forestall.
>
> In the first place, I am not claiming that scientific theories are formulated in such a way that they could not possibly be multiply realized. I am claiming only that it is reasonable to hope that a good theory will not in fact be multiply realized.
>
> In the second place, I am not claiming that there is only one way in which a given theory *could* be realized; just that we can reasonably hope that there's only one way in which it *is* realized.[10]

Further below I'll return to the question of how convinced we should be of all this. For now, let us just note that the New theoretical terms are explicitly definable in the now reformulated theory, though with the use of higher-order quantification. There is then no longer a question about how to understand them: those definitions spell out their meaning completely.

6. Anti-Nominalism

There was a hidden flaw in the above proposal which was brought to light by Putnam's argument.

Let us consider a scientific theory T, and let R′ be the open sentence resulting from replacement of all the New terms in T by variables of appropriate category. Let R be the existential closure of R′—that is to say, the Ramsey sentence of T.

(1) Assuming consistency throughout, there will be a model in which R′ is true on some assignment of values of the variables.
(2) If T has an infinite model, then R′ has many non-isomorphic such realizations, each of which yields a model in which R is true.

From these it follows that if T is consistent and has an infinite model, then Lewis' proposed uniqueness postulate is false. The model theoretic facts appealed to here are basically those used in Putnam's model-theoretic argument—in this sense, Putnam's argument could be taken as a direct objection to Lewis' theory of science.

So Putnam had indeed devised a bomb, as Lewis said; could it possibly be defused? The answer is Yes, of course; for whatever logic can do, a little more logic can undo. All the weight in the argument I have just sketched is really borne by the very innocuous idea that

there is something x such that Fx

is true on an interpretation (including a value assignment to variables) exactly if there is some (other) interpretation differing from it at most in what gets assigned to *x*, and which satisfies *Fx*. The question whether or not there is, is a mathematical question of form: does there exist a function with such and such properties? And here, at least as model theory is taught in most college courses, *any and all functions* assigning values of appropriate type to the variables will be admissible. To variables that take the place of names they can assign any member of the domain, to variables taking the place of predicates of degree one they can assign any subset of the domain, and so forth. Without this leeway, this unbounded admissibility, neither Putnam's own argument nor its adaptation just sketched could go through.

Thus Lewis' problem. To put it simply: existential qualification over properties will not yield any substantial claims, if arbitrary sets can be the sets of instances of properties.[11] Hence also Lewis' solution: in our reading of the Ramsey sentence, and of his uniqueness postulate for the theory, we must take the variables (and corresponding quantifiers) to have restricted ranges. The most important here (as our overly simple 'water theory' illustrates) are the predicate variables. Those variables must be read as ranging not over arbitrary subsets of the domain, but only over *certain* sets—call them the "natural sets" or "natural

classes". Assuming that existential generalization is still a valid logical move, this means also that the predicates in the scientific theory under construal are to be assumed to have natural classes as their extensions.

What content could there be to this distinction between natural and arbitrary classes? It has the authority of a venerable philosophical tradition. Depending on your specific metaphysical persuasion, you can take natural classes to be the extensions of real properties, or of universals. The niceties of such elaborations are not here to the point, but do mean that *anti-nominalism* was already waiting in the wings for just such a rescue operation.

7. Clashing Views of Science

The moving force in the dialectic just now presented is the drive to preserve Lewis's view of science in the face of Putnam's paradox. Assuming that aim, we found a compelling argument for anti-nominalism. But could we demur?

There are alternate views of science, and if we cannot very well expect a decisive battle between them, it may be instructive simply to bring them to mind. To begin, both Lewis and Putnam pose their problems in the discourse of logical positivism (even if they are perhaps doing this partly in dialectical concession). Both depict a scientific theory as a set of sentences formulated in a specific language. Both write in terms of a two-fold division of that language's non-logical vocabulary. Lewis explicitly distances himself from the positivist dogma that identified Old terms with pure (theoretically hygienic) observation predicates. The problem of how to understand theoretical terms, given only that we understand observation predicates, has been transformed by time-indexing, so to speak: how shall we understand, or try to explain, the meaning of New terms, introduced for a frankly theoretical purpose, and for which no explicit definitions are made available?

This is not very far from the positivist *problématique*. We may honor its desire to understand without sharing all its presuppositions. Specifically, we need not agree that the best possible solution is to recast theories in such a way that the New terms do after all receive explicit definitions, or implicit definitions, or partial definitions through reduction sentences, or the like. Recall Lewis' own dictum about the understanding of Old terms:

> by 'understand' I mean 'understand'—not 'know how to analyze'

Won't this sauce for the goose go very well with the gander too? We do come to understand the New terms introduced in science—understand and use, *not* know how to analyze or define!

The alternate view to explore is therefore exactly that we can acquire a richer (truly richer!) language than we had before. We can acquire new resources that allow us to say things that we truly did not have the resources to say. It may be mysterious how we can do this, though it seems that every child does; indeed,

myths aside, humanity must have fashioned its language, like its other tools, itself. The problem of how we learn language is not a philosophical problem. How new terms are introduced, how their use becomes stable, and how that use is communicated from person to person, is a real problem but not a philosophical one. It is an empirical problem, to be investigated scientifically, and not bedabbled in the metaphysician's armchair.

The second clash of views I wish to have in the open is still about language, while the third is not. But I can bring both to the fore by asking what the history of science would ideally be like, seen through Lewis' representation of scientific theories. Let me emphasize that in doing so I am clearly stepping outside Lewis' project in "How to define theoretical terms", which considered only what happens *at the time* when theoretical terms are newly introduced. But it seems to me that we must also ask how a construal of what happens at particular moments can extend to a view of how science and language evolve over time.

First of all, on Lewis' view there is in scientific change no change in the understanding or use of the Old terms. This is exactly what Feyerabend calls the 'stability thesis' concerning the language of science: that Old terms retain their old meaning when new theories are introduced. Feyerabend's arguments against this thesis are not overly dependent on any particular view of meaning. In the scientific examples he displays we see changes in reference, in extension, in use, in connotation, and in logical implications of the sentences in which the Old terms appear. On the face of it, at least, science cannot rest on a complete understanding of the Old terms, retained throughout theory change. Without that assumption, however, the problem that Ramsey, Carnap, and later Lewis set themselves becomes moot.

Secondly, let us imagine that science develops under the best possible circumstances of disinterested intelligence. We can imagine that scientists introducing new theories do so naively, to begin, with New terms introduced for the purpose. Then they cast the theory into 'canonical' form using Lewis' recipe for transforming it (roughly speaking, into its Ramsey sentence plus a uniqueness postulate), relative to which the New terms are explicitly definable. Let us, for brevity, write T* for the result of transforming naive theory T in accordance with Lewis' recipe.

This apparently simple story actually harbors two ambiguities. The first thing we notice is the sensitivity to the time element. Suppose that new vocabulary together with new theory T is introduced at time t, and that at later time t′ a new 'empirical' postulate is added (call it A) which uses only the terms already Old at t. Now what is the meaning of the New terms? Their explicit definition relative to T* is not the same as that relative to (T+A)*. On the other hand, there was no introduction of New terms, hence no occasion to employ Lewis' recipe.

We may retort that it really does not matter: shown to be explicitly definable, the New terms were of course shown to be dispensable, epiphenomenal, irrelevant. Fine. But if the question was: *what do the New terms employed in science*

mean? we do not have an unequivocal answer. That this does not matter will at best confirm our feeling of unreality in such a discussion of science.

Besides, the diachronic ambiguity is accompanied by a synchronic one. Recall that when I illustrated the transformation with the example of a small theory of water, the theory was too small—it would be wrong to understand it as carrying a uniqueness postulate. So those imaginary scientists must be counseled to cast into canonical form only 'large enough' parts of their science. The transformation is not to be carried out on the innovative new theory that introduced the New terms, but on its sum with a surrounding shell of background theories. How large a shell? We will not get the same definitions of New terms if we choose larger shells.

This reminds us also of a more obvious problem: if the asserted uniqueness is unreasonable for 'small' theories, won't the danger of choosing too small remain unless we take in the whole of science? (And would it not be better still to wait a while till this edifice has been improved and unified? No, that really puts the idea at the end of the rainbow!) Once more, the simple story, though predicated on an explicit idealization to impart precision, turns out to be ambiguous.

Finally, let us look back through this ideal history to the beginning of time. Each time New terms were introduced, they were effectively eliminated, though presumably retained for vulgar use in practice. So: all the terms that *really* belong to the language of science at any stage of its history were there all along. They are the 'absolute' Old terms, the ones that were already Old when science began. Perhaps there were false starts along the way, when scientists accepted hypotheses (always expressible, as we now see, entirely in Old terms) that had to be given up later on. Erasing those missteps from the official record, and ignoring the uniqueness postulates for a moment, we see a steady accumulation of theory, within the Old language, as science learns more and more about what the world is like. Accompanying this cumulative growth of information, however, is a steady stream of uniqueness postulates, each weaker than the one before.[12] Thus any appearance of scientific revolutions is merely appearance. The official history of science so construed is quite different from what many now—following Kuhn, Feyerabend, and a host of others—take the real history of science to be like.

Being at odds with the currently received view is generally a virtue; it certainly does not count as an objection. But the received view is not egregious, by definition so to speak; so the main point stands. On the face of it, we may indeed demur from the dialectical aim that led from Putnam's paradox to anti-nominalism.

8. Taking Stock: Where Do We Stand Now?

In the first three sections of this paper I proposed a reading of Putnam's argument that dissolves the threatened paradox by closer attention to the use of "true" in our own language for sentences of our own language. The next four sections analyzed Lewis's discussion to see where the support for anti-nominalism

can be resisted. If my analysis is correct then recourse to anti-nominalism is, in this instance anyway, the solution to what is—from an alternative tenable point of view—an ill-posed problem. Where do we stand now, and how much of the paradox remains to haunt us?

In the middle part of this paper I analyzed how Lewis turned Putnam's argument into support for a more articulate metaphysical position. Lewis used Putnam's argument to refute views of language clearly in competition with views Lewis advocates. If we can be be brought to a choice among these competitors, we shall accordingly have to opt for Lewis'. Two lines of argument can bring us to such a choice point. The first begins with a putatively mandatory philosophical question ("How does language as a whole get its meaning?", "How is reference fixed?" to use two slogan formulations) which apparently has only two sorts of feasible answer, global descriptivism and anti-nominalism. Putnam's model-theoretic argument refutes the former.

The second line of argument begins with an seemingly more modest but more obviously mandatory philosophical question: what is the meaning of a theory stated by means of newly introduced theoretical terms? Lewis had earlier proposed a construal of scientific theories on which the newly introduced vocabulary is definable in terms of the old, prior vocabulary. This proposal was a concrete version of what he calls "local descriptivism" in his paper on Putnam. But it also confronts us with a dilemma. For if the proposal is coupled with an understanding that all interpretations, all assignments of classes to predicates as their extensions, are in principle equally eligible, then Lewis' construal makes almost every scientific theory obviously false—a point establishable as corollary to Putnam's model-theoretic argument. Thus the benefits of this proposal are not reaped unless a suitable restriction is imposed: enter anti-nominalism, once more.

I have objected to both lines of argument in more or less the same way. The alternative posed is based on assumptions which we need not (and, in my opinion, should not) accept. The putatively mandatory philosophical questions are not mandatory in themselves, but only relative to those assumptions. Can I support my contention that we should not accept those assumptions, that we should not allow ourselves to be led into those dilemmas and trilemmas, that the questions are not after all mandatory philosophical questions? That is the first remaining issue.

On my reading of Putnam's model-theoretic argument, the paradox dissolves. What remains is a striking reductio of a certain view of language, which we can independently verify to be inadequate. Perhaps that was just what Putnam intended; perhaps the view of language found wanting is implied by that correspondence theory of truth which Putnam locates at the heart of metaphysical realism. I would like to think so; but authorial intent is notoriously indiscernible; the text has broken its moorings and must in any case be dealt with on its own terms. More salient at this point is the question what assumptions were involved in my own reading, and whether they can be further supported if pressed. That is the second issue.

A moment's reflection shows that the two issues which remain are very closely related. They require us to step back, to some extent, from the immediate intricacies of Putnam's argument to a wider perspective on language.

9. In the First Person: Problems and Pseudo-Problems

The crucial move in Putnam's 'model theoretic' argument occurs at the end, where "true" is equated with "true under some interpretation".[13] This equation is incorrect as it stands, since it equates truth with satisfiability. While that observation suffices to block the conclusion, in preliminary fashion at least, it leaves the innuendo standing. To escape the argument's spell, we need to distinguish carefully between the entangling pseudo-problems and the real problems they resemble. This has much to do with something that is not expressible at the level of semantics: which language is *our* language, and how our language in use relates to languages under discussion.

Truth in our own language. For texts in our own language, attribution of truth is not elliptic, and 'true' does not mean 'true relative to some interpretation'. But while "true" is then not elliptic, it's use is indexical—tacitly indexical. The tacit indexical reference is to our own language. Criteria for proper understanding of our own language express themselves in *pragmatic tautologies*. Consider the sentences:

"cat" denotes cats.
"Paul is a cat" is true if and only if Paul is a cat.
"gargel" denotes gargels.[14]

The third sentence I need not assert or endorse; I do not so much deny it as reject it, thereby signalling that I do not count "gargel" part of my vocabulary in use. But the first and second sentences are paradigmatic examples of pragmatic tautologies in my language. They are undeniable by me, exactly because I acknowledge "cat" to be a word in my language. The semantic content, however, of these (to me undeniable) assertions are not necessary propositions, and most certainly not tautologies in the sense of semantics. If our language had developed differently in a certain way then "cat" would have denoted gnats, rats, or bats. Under such circumstances, uses of "cat" would not have been acts referring to cats, and "Paul is a cat" would have been used to state that Paul is (not a cat but) a gnat, rat, or bat. Pragmatic tautologies (for me) are sentences of my own language which state something that could indeed be (or could have been) false but which I cannot coherently deny.[15]

Such pragmatic tautologies are statements which are undeniable to us in whose language they are formulated, while not expressing anything necessary. What they say could be false; but we cannot coherently assert anything contrary to them. Moore's paradox is the most familiar example. I cannot assert "It is snowing in Peking and I do not believe that it is snowing in Peking", even though

there are many times when both conjuncts are true simultaneously. But " 'cats' denotes cats" and " 'Snow is white' is true if and only if snow is white" are also good examples of philosophical interest. (Perhaps Descartes' "I exist" and Putnam's "I am not a brain in a vat" are too; but let us leave such more recherché examples aside.) Our proper avowal of such statements must be accompaniable, somehow, with an ability to admit the possible falsity of what they say. We must even be able to acknowledge the possibility of a radical failure in our own acquisition of the language we speak—a way to acknowledge that possibility in the words of that very language!—without reducing ourselves to incoherence.[16] However that may be, we must be careful not to confuse the undeniability of a pragmatic tautology with certainty of its content. This point has several applications.

How is reference fixed? That question has a presupposition, conveyed unfortunately only by metaphor. For of course we don't use glue or even hammer and nails to attach words to things, nor does nature glue itself onto our words—even the word "attach" merely continues the same metaphor. So what is the problem? Abstractly stated it is this: each of our predicates has an extension, and might have had a different extension. But unless they have the right extension, we can't use our language to frame genuine, non-trivial empirical statements or theories. So, under what conditions do they have, or acquire, the right extensions?

This abstract statement and question have the form of intelligible, non-trivial expressions. But form is not enough. Let us see what happens when we get down to brass tacks with them. Take the word "green", which we use in making statements about parts of the world well beyond our ken. Now, what is the worry when we worry that this word might not have the right extension? The only answer I can come up with here is:

> the worry that there are lots of green things out there which aren't in the extension of "green" and/or things that are not green yet are in that extension.

But what sense do I make if I say to myself:

> There are green things which are not in the extension of "green".
> There are some things x such that x is green but "is green" is not true of x.

If I say this sort of thing I do not make sense. I may convey through this utterance either that I have no grasp of the philosophical jargon ("extension", "is true of"), or that I do not acknowledge the words (e.g. "green") in that sentence as belonging to my vocabulary. The worry that there might be green things out there not denoted by "green"—or cats not denoted by "cat"—is a pseudo problem.

How does translation work? What made it seem that we had hold of a real problem when we stated it in the abstract? Two things, not unrelated to each other. The first is that nonsense can have the form of sense. Many analogues to such 'form-generated' puzzles are known: the fear that all fears are unfounded and are

in fact symptoms of paranoia, the insistent request to keep something for me until I give it back to you, and so forth. The second is that there real questions very closely related to the pseudo question, and easily confused with it. In our particular case, the nearness is that of seemingly uncontroversial translation. Think of a Dutch speaker, Piet, who is investigating an alien language, English, and asks:

(1) Is "cat" een woord voor katten?

This is undoubtedly a non-trivial empirical question.[17] Now translate it into English:

(2) Is "cat" a word for cats?

Is that a non-trivial empirical question? Does it have exactly the same cognitive, epistemological status as the original? We note first of all that (2) is not an empirical question for his English informant Jane. If she says Yes, she is only uttering a pragmatic tautology, something she cannot deny (if she acknowledges all the words in (2) to belong to her language). But secondly, (2) does not have the same status as (1) even for Piet! For it is perfectly possible for him to have come so far in his study of English that he recognizes (2) for what it is, while having no idea what the English word "cat" denotes.

Piet's real problem is not a philosophical problem. He can solve it only by empirical methods.[18] If we translate it for him into philosophical jargon, we don't get him one step further. There is a philosophical problem highlighted by this example, if you care to note it: what is translation? For how can (2) be a perfect translation of (1) if they do not have the same status for Piet, and (2) does not have the status for Jane that (1) has for Piet? But if this is not a case of perfect translation, what is wrong with it?

Translation's intralinguistic companion. Since it is a pragmatic tautology that cats are what "cat" denotes, why can't I replace the one expression by the other everywhere? Yet to know that "cat" denotes the things denoted by "cat" is not at all to know that "cat" denotes cats. Moreover, that the latter (the Yes answer to (2) above) is a pragmatic tautology cohabits in us with the insight that "cat" need not have denoted cats, but could well—given some alternative evolution of our language—have ended up denoting gnats, rats, or bats. It has been famously suggested about all this is that we must distinguish the a priori from the necessary. But this may not have ended all our puzzlement.

The worry that the extension of "cat" is not fixed is the worry that the answer to (2) may be No, though we cannot answer it with "No". But that worry, which makes no sense, is only the confused echo of the real worry that empirical question (1) has answer No. That there are both real empirical questions and interesting philosophical problems in the neighborhood does not mean that the 'how is reference fixed?' problem is anything but a pseudo problem. Obviously we don't do any fixing, and whether nature is doing any gluing is just beside the point.

Understanding our own language does not reside in having an interpretation for it. Using one part of our language we can interpret another part. In this way I can give non-trivial true information about my language: "dog" does not denote cats, "woodchuck" denotes groundhogs. This is exactly analogous to giving information about other languages: the Dutch word "kat" denotes cats. But in the context in which I do so, I use and rely on (part of) my own language, and in that context, the same questions do not arise for the part on which I rely. Trying to press the project for my language as a whole, I can end only in pragmatic tautologies.

If nature does not fix or restrict reference, then we must be doing it, by our intentions, practice, or by other constraints we can impose on the use and understanding of our language—how is that possible? This question has exactly the presuppositions of the preceding one about fixing reference, and does not arise at all if those presuppositions are rejected. (Sometimes the alternative to anti-nominalism pointed to in this confused question is even called "anti-realism", as if imputing metaphysical powers to persons is not realist metaphysics!) Trying to complete an interpretation for my language as a whole, in some independent, informative, non-tautological way, can only reduce us to absurdity. For interpreting is an activity involving use of and reliance on my own language and inconceivable without it. "Relying on my own language" does not mean "assuming that we have grasped a complete interpretation of it", of course, for that would make the notion of reliance incoherent as well. To think so is just to keep bringing back in that wrong philosophical view of understanding language which fell apart when we first examined it.

The flybottle is difficult to escape. Let me try to say the same thing in still another way. As Putnam has also pointed out, things simply don't get any better with reification of metaphysical intermediates. Suppose I try to avoid the absurdity and triviality of grasping a complete interpretation of my own language by intermediate stages, such as the following:

> "cat" connotes cathood.
> Cathood has as instances all and only those things which are cats.
> Words denote exactly the instances of what they connote.

I will have enlarged my language, but the same point will apply to the larger language, and nothing will be gained.

Suppose alternatively that I say: the denotations of my words are fixed by us, in some ways which we cannot fully express in our own language, but which supply the external constraints needed to supplement the internal constraints on reference. Obviously I must be assuming that the internal constraints do not cut down the range of available interpretations sufficiently to avoid triviality, *and that something must be doing so.* This assumption turns a very ordinary sort of problem into a philosophical pitfall. That some people use "cat" and others "gatto" to denote cats is as ordinary a fact as that some use hammers and others use crowbars to pull out nails. But *sub specie* that assumption we must, if pressed, come up with ideas like: we have ways to grasp cathood and to connote it with one

of our words, or to single out natural classes and denote them, or some such completion. Or we can amend: it is not that we have ways of doing such things, but that kind Nature, Providence, or Deity has arranged it for us, programmed us to speak in a way that latches onto inexpressible features of reality, and so forth. We have a language blessed among all others by this special relation to reality... or at least, we can *hope* that we do, and proceed with the *faith* that we succeed in meeting those 'external' constraints on reference, in the *charity* of assuming that others do as well... .

There is no help for these impasses except philosophical therapy. Putnam skillfully purveyed, in the course of his reductio argument, the picture of language according to which to understand or have (!) a language is to know its syntax and to grasp an interpretation of that syntax. This picture is nonsensical, as comes to light as soon as we ask: in what language is this grasp expressed, in what language do we describe this interpretation that we grasp? Now I imagine the metaphysically inclined might even think of postulating a non-verbal, or inexpressible, grasp of some interpretation of our own syntax, to give meaning to that syntax in our mouths. But this postulate would hover between the trivial assertion that we speak meaningfully by uttering words in our language, and a bit of armchair psychology, pseudo science to fill the perceived gap, to give the appearance of an explanation of how humans are able to speak meaningfully. I say: let's be content with the trivial assertion acknowledged as such and leave scientific accounts of psychological phenomena to empirical science.

The familiar limits of relativism. My dissolution of Putnam's paradox clearly hinges on our philosophical resources being adequate to their job, and hence on the coherent elaboration of what I have just said. But I am heartened by the resemblance of those further problems to others that we have equal reason to dissolve. Consider this view about values, whether of morality, prudence, or some more restricted value domain: what is good is what meets our standards. The word "our" is to be taken seriously, its indexical character crucial to meeting standard objections to value 'relativism'.[19] The idea is that the Romans could also have correctly expressed an important truth by using those very same words. But we add that there were certain moral insights that the Romans, indeed humanity, did not yet have concerning slavery (or concerning the weighing of probabilities in prudence, etc.). That is quite consistent, since of course in this addition, the tacit reference is once more to *our own* standards to ground the evaluation. And now the puzzle comes, when we add: but of course it is possible that we are with respect to some issue in the same position as the Romans were with respect to slavery. What do we mean? That there is something which is good, unbeknownst to us, although it does not meet our standards? That would seem to scuttle the view we are trying to elaborate.[20] Whatever the merits of such a view, it should not be scuttled by logic; and the logical pitfalls by which it is beset resemble just the ones we have discussed here.

As I said to begin, Putnam's paper is very rich. The paper richly illustrates Poincaré's quip about logicism: [meta]logic is not sterile, it engenders paradoxes. If fortune will, the engendered paradoxes will lead us to new insight.

Notes

*An earlier version of this paper was circulated in ms. as "Putnam's Paradox Revamped". It is companion to two others: my "Structure and Perspective: Philosophical Perplexity and Paradox", (pp. 511–530 in M.L. Dalla Chiara et al. (eds.) *Logic and Scientific Methods*. Dordrecht: Kluwer, 1997.) and "Elgin on Lewis' Putnam's Paradox" *Journal of Philosophy* 94 (1997), 85–93. I wish to acknowledge my great debt to the writings of David Lewis and Catherine Elgin, noted below. Discussions and correspondence both with them and with Igor Douven have been extremely helpful. Jenann Ismael, Mary Kate McGowan, Elijah Milgram, Chad Mohler, Laurie Paul, Gideon Rosen, and Jill Sigman helped me with critical comments.

1. "Realism and reason", Presidential Address to the Eastern Division of the American Philosophical Association, December 1976; reprinted in his *Meaning and the Moral Sciences*, 1978, pp. 123-140.
2. David K. Lewis "Putnam's Paradox", *Australasian Journal of Philosophy* 62 (1984), 221-236.
3. For extensive discussion, see D. Lewis, "New work for a theory of universals", *Australasian Journal of Philosophy* 61 (1983), 343-377, and e.g. my *Laws and Symmetry* (Oxford 1989), Ch. 3, section 5, and Elgin (see below). As Lewis notes, Putnam explicitly rejects Lewis' response to his model theoretic argument (cf. H. Putnam, *Reason, Truth, and History* (Cambridge 1981), p. 53). As Lewis also notes, the anti-nominalist solution was first discussed (but not advocated) by Gary Merrill, "The model-theoretic argument against realism", *Philosophy of Science* 47 (1980), 69-81.
4. Catherine Z. Elgin, "Unnatural science", *Journal of Philosophy* 92 (1995), 289-302; see also my "Elgin on Lewis' Putnam's Paradox", *Journal of Philosophy*, 94 (1997), 85–93.
5. Putnam himself indicates such constraints, cataloguing them as prediction of observation sentences, 'operational constraints', and internal theoretical virtues such as simplicity. The burden of the text (perhaps signalled explicitly by the "(as far as we can tell)") is surely (?) that these constraints do not suffice to fix the extensions of the L vocabulary in THE WORLD sufficiently so as to prevent the two-stage interpretation via an arbitrary model M of T_l from being admissible.
6. Catherine Elgin explores this response (*op. cit.*).
7. In this approach to the 'true' import of the model theoretic argument, I am attempting to follow directions in Putnam's own discussions of representation. Imagine for example that an ant makes a track in the Sahara desert which to our eyes would look like the word Coca-Cola in cursive script. Has the ant written "Coca-Cola"? We can similarly imagine that in extraterrestrial travel we might someday come across a rock formation that to our eyes exactly resembles Abraham Lincoln. Is that a bust of Abraham Lincoln? They certainly become representations of a requisite sort if we put them to representational use. The idea that independently of that use they are already representations (as opposed to objects suitable for use as representations), can only be based on too simpleminded an idea of representation. Compare Putnam's 'internal realist' account—in the very paper that presents the model theoretic argument—of why language-using contributes to our success in practical affairs: "[it] is not that language mirrors the world but that *speakers* mirror the world—i.e. their environment—in the sense of *constructing a symbolic representation of that environment.*" (Putnam 1978, p. 123).
8. I can hear a little voice saying "But if the argument is general, what can be wrong with instantiating its conclusion to any language, whether it is ours or not?" That little voice is too much in the grip of the standardly taught logic, ignoring (as courses and text

books often do) logic's self-imposed limitations. To give a much simpler example, what could be wrong with the rule to infer B from (A&B)? But you cannot apply that rule if (A&B) is "Snow is white and *this* is a conjunction".

9. David K. Lewis, "How to define theoretical terms", *Journal of Philosophy* 67 (1970), 427-446; reprinted as Chapter Six of his *Philosophical Papers* (Oxford, 1986); citations are from p. 79, 80. These passages are part of Lewis's response to John Winnie's argument that every theory must have multiple realizations; Winnie's argument relies on variation in the extensions of the Old terms. See J. Winnie, "The Implicit Definition of Theoretical Terms", *British Journal for the Philosophy of Science*, 18 (1967); 223-229.
10. *Ibid.* pp. 83-84. The phrasing makes one a little uneasy; what is the proposal exactly? The original proposal was that replacement of a theory by its Ramsey sentence incurs no real loss. If so we can, for all purposes of philosophical reflection, think of scientists as introducing those Ramsey sentences plus some *façons de parler*. Is Lewis asserting this of his amended version, or is he not? Is he saying that we can construe the scientist as asserting the unique realization of the structural pattern, or does he mean that we should construe them as asserting the Ramsey sentence, while hoping that they will come up with ones that are not multiply realized, in just the way that we hope they will come up with ones that imply true predictions? I will assume the former.
11. As Lewis also notes in "Putnam's Paradox", and as had been pointed out by W. Demopoulos and M. Friedman ("Critical Notice: Bertrand Russell's *The Analysis of Matter*: its Historical Context and Contemporary Interest", *Philosophy of Science* 52 (1985), 621-639), this was exactly the criticism that had earlier been given by M.H.A. Newman of Russell's structuralism, in his "Mr. Russell's 'Causal Theory of Perception' " *Mind* 37 (1928), 137-148.
12. Weaker, because when the Ramsey sentence has content added inside the scope of its initial existential quantifiers, the assertion that the pattern of properties and relations it now describes is uniquely instantiated, is weaker. We must assume that Lewis means the earlier, stronger uniqueness postulates to be discarded in favor of he later, weaker ones.
13. I have examined this move more closely in "Elgin on Lewis' Putnam's Paradox".
14. I shall use the sentences " 'cat' denotes cats", "the extension of 'cat' is the set of cats", and " 'cat' is a word for cats" interchangeably here. As Catherine Elgin pointed out to me, the more normal use of "denote" implies the existence of what is denoted—in this more usual sense, we would say that "cat" denotes cats if there are any.
15. The sentence " 'Paul is a cat' is true if and only if Paul is a cat", *as understood by me*, is false in that other envisaged situation. We are here envisaging a possible situation in which people use "cat" as a word for e.g. gnats, but would of course themselves also assert the sentence (= wellformed bit of syntax) " 'cat' is a word for cats". When we describe that situation, all the words we use have their normal everyday meaning. But some of those words we also mention; those we mention, we say, have a different meaning in that other situation. Accordingly, the general format for discussion of truth becomes

    A, as understood in x, is true in y.

    There exists a hybrid of semantics and pragmatics called "two-dimensional semantics", in which such distinctions are represented in a certain way. I used to be enam-

ored of it more than I am now. See my "The only necessity is verbal necessity" (*Journal of Philosophy* 74 (1977), pp. 71-85) for, among other applications, a sort of semantic representation of the sentences I would now classify as pragmatic tautologies.

16. This problem has so far been studied mainly in connection with another controversial candidate for the status of pragmatic tautology: the Reflection Principle for subjective probability ("It seems very likely to me that it will rain tomorrow, on the supposition that tomorrow morning it will seem very likely to me that it is going to rain"). See my "Belief and the Will" (*Journal of Philosophy* 81 (1984), pp. 235-256) and "Belief and the problem of Ulysses and the Sirens" (*Philosophical Studies* 77 (1995), 7-37).
17. For the sake of example I am here pretending that Dutch and English are two separate languages *in actu*. I usually think of one's language as everything one has learned to speak, and of natural language as consisting in all the resources we have for speaking and writing. See my "The World We Speak Of, and the Language We Live In", pp. 213-221 in *Philosophy and Culture: Proc. of the XVII-th World Congress of Philosophy*, Montreal 1983 (Montreal: Editions du Beffroi, 1986).
18. As Catherine Elgin reminded me, there is another empirical problem in this neighborhood, which may add to the confusion. Suppose we regiment our language by agreeing to some explicit criterion for what is really green. (It might be, e.g. reflecting light of a certain wavelength.) Then we may worry that all our paradigm examples in the past, for which we introduced the word "green" in the first place, were actually things which only looked green to us under current conditions but do not meet that criterion. Such doubts too are handled appropriately only by empirical research; they are not sceptical doubts.
19. I call such views of morality "Cosa Nostra views"; they were recently explored by my (then) colleagues Michael Smith, Mark Johnston, and David Lewis. Such views seem to me to be at least very apt for more practical forms of evaluation (prudence, 'instrumental rationality' and the like); their tenability for ethics we may leave aside here.
20. Wilfrid Sellars framed this problem for the emotive theory of ethics, but to highlight difficulties with the substitution interpretation of quantifiers. However analytic we get, however abstruse or scholastic our philosophical preoccupations, wrenching existential problems lie in wait for us everywhere, and all philosophical problems are connected...or so it seems... .