Australasian Journal of Philosophy
Vol. 62, No. 3; September 1984

PUTNAM'S PARADOX

David Lewis

Introduction. Hilary Putnam has devised a bomb that threatens to devastate the realist philosophy we know and love.[1] He explains how he has learned to stop worrying and love the bomb. He welcomes the new order that it would bring. (RT&H, Preface) But we who still live in the target area do not agree. The bomb must be banned.

Putnam's thesis (the bomb) is that, in virtue of considerations from the theory of reference, it makes no sense to suppose that an empirically ideal theory, as verified as can be, might nevertheless be false because the world is not the way the theory says it is. The reason given is, roughly, that there is no semantic glue to stick our words onto their referents, and so reference is very much up for grabs; but there is one force constraining reference, and that is our intention to refer in such a way that we come out right; and there is no countervailing force; and the world, no matter what it is like (almost), will afford *some* scheme of reference that makes us come out right; so how can we fail to come out right?[2]

Putnam's thesis is incredible. We are in the presence of paradox, as surely as when we meet the man who offers us a proof that there are no people, and in particular that he himself does not exist.[3] It is out of the question to follow the argument where it leads. We know in advance that there is something wrong, and the challenge is to find out where. If the paradox-monger is good at his work, we stand to learn something; and indeed, I think that Putnam's paradox affords an important lesson.

In the first half of the paper I shall give my account of what I take to be the core of Putnam's argument, and I shall say how I think it fails. In the second half of the paper, I shall raise some questions about aspects of Putnam's presentation that puzzle me.

[1] Hilary Putnam, 'Realism and Reason', *Proceedings of the American Philosophical Association* 50 (1977) pp. 483-498, reprinted in Putnam, *Meaning and the Moral Sciences* (Routledge and Kegan Paul, 1978), henceforth 'R&R', cited with page numbers from *Meaning and the Moral Sciences;* 'Models and Reality', *Journal of Symbolic Logic* 45 (1980) pp. 464-482, henceforth 'M&R'; and *Reason, Truth and History* (Cambridge University Press, 1981), henceforth 'RT&H'.

[2] Compare the malicious joke: 'Mr. Z claims to have found a counterexample to my theory. But he has misunderstood me, he has not interpreted my words as I intended. For I intended that there be no counterexamples.'

[3] Peter Unger, 'Why There Are No People', *Midwest Studies in Philosophy* 4 (1979), pp. 177-222; and 'I Do Not Exist', in G. F. Macdonald, ed., *Perception and Identity* (Macmillan, 1979).

Three caveats. (1) I warn the reader that I am not sure how well I understand Putnam.[4] Sometimes, different things he says seem to point in different directions. What is more, I shall state what I take to be his argument in my own way. I hope the line of argument I discuss is Putnam's, even if rather freely paraphrased. But whether it is his or not, I think it worthy of attention. (2) I shall acquiesce in Putnam's linguistic turn: I shall discuss the semantic interpretation of language rather than the assisgnment of content to attitudes, thus ignoring the possibility that the latter settles the former. It would be better, I think, to start with the attitudes and go on to language. But I think that would relocate, rather than avoid, the problem; wherefore I may as well discuss it on Putnam's own terms.[5] (3) I shall ignore the complex details of model-theoretic semantics for natural language. I suppose that a proper treatment would require interpretations in which the semantic values are elaborate set-theoretic constructions.[6] But I shall acquiesce in Putnam's supposition that we can get by with model theory in its 'basic form' (R&R, p. 124): we have a domain of 'parts of the world', things which may serve as referents for singular terms, and classes of which may serve as referents for general terms. Such a supposition might matter, if model-theoretic results were as important to Putnam's argument as he suggests. But I load the dice in Putnam's favour, if at all; so I play fair.

Global Descriptivism Refuted. We are familiar with the idea of a description theory of reference—for short, *descriptivism*—and, especially, with a local form thereof. Suppose, *pace* Putnam, that somehow we already have an extensive language with fairly determinate reference. Then we may add new language to the old, a little at a time, by introducing undefined terms in our theorising. Thereby we associate clusters of old-language descriptions with our new terms; and thereby, if the world cooperates, we bestow reference on the new terms. 'Jack the Ripper did this, that, and the other' says the detective; his point is in part to hypothesise that there is someone who did this, that, and the other, and in part to stipulate that the one who did, if such there be, is to become the referent of 'Jack the Ripper'. The new term 'Jack the Ripper' is to acquire the referent, if any, of the old-language description 'the one who did this, that, and the other'. The intended interpretation of the augmented language is to be an extension of the old interpretation of the old language, if such there be, that makes the new Jack-the-Ripper theory come true.

[4] I find it especially hard to make RT&H mesh with R&R and M&R, but I do think they are suppposed to mesh. The third full paragraph of RT&H, p. 7, indicates a connection. Also, RT&H was in draft before Putnam read Goodman's *Ways of Worldmaking* (see RT&H, p. xii); the latter was published in 1978, and might have been available in manuscript to a sympathetic colleague earlier than that; so RT&H is more nearly simultaneous with R&R and M&R than their publication dates would suggest.

[5] For a discussion of the 'relocated' problem and its solution, see the final section of my 'New Work for a Theory of Universals', *Australasian Journal of Philosophy* 61 (1983), pp. 343-377.

[6] More or less as in my 'General Semantics', *Synthese* 22 (1970), pp. 18-67.

Seven points should be noted. (1) There may or may not be rigidification. If there is, that will avoid confusion between people who have attached the same term to the same referent by means of different descriptions. For nothing will be true as one person means it but false as the other means it, not even when the term appears in modal contexts.[7] (2) The term-introducing descriptive theory may be egocentric (for instance, it might include 'Water is abundant on this planet'); (3) it may make reference (in old language) to word tokens or thought tokens; and (4) it may involve relations of causal acquaintance. Taking points (2)-(4) together, we note for instance that 'Beech trees are the causal source in such-and-such way of tokens in my speech and thought of "beech tree"' might be part of the bit of descriptive theory that, for me, attaches a referent to 'beech tree'; and so, *mutatis mutandis,* with 'elm tree'. (5) The description needn't fit perfectly. 'Jack the Ripper' might take as referent the one who comes closest to doing this, that, and the other, if no better candidate is available. The intended interpretation of the augmented language, then, is to be that extension of the old interpretation that comes as close as can be to making the new Jack-the-Ripper theory come true. (6) There might be two candidates that both fit perfectly; more likely, there might be two imperfect candidates with little to choose between them and no stronger candidate to beat them both. If so, we end up with indeterminate reference (in addition to whatever results from indeterminacy of the old interpretation of the old language): the new term refers equally to both candidates. Hartry Field's example of Newtonian 'mass' illustrates this possibility.[8] Note well that this is moderate indeterminacy, in which the rival interpretations have much in common; it is not the radical indeterminacy that leads to Putnam's paradox. I take it that the existence of moderate indeterminacy is not to be denied. Finally, and most important for what follows, (7) it may happen that new terms acquire their referents by description not singly but in families. Suppose that our detective hypothesised that the murders were the joint work of a couple: Jack the Ripper and Jill the Slasher, as he chose to call them. 'Jack did this,' he says, 'Jill did that and the other, and Jack and Jill are related thus'. Then, if the world provides suitable candidates, 'Jack' and 'Jill' gain referents together. The intended interpretation of the doubly augmented language is to be an extension of the old interpretation of the old language, if such there be, that makes the new Jack-and-Jill theory come true.

Description theories of reference are supposed to have been well and truly refuted. I think not: we have learnt enough from our attackers to withstand their attacks. I think that a descriptivism that takes to heart the seven points just listed is still tenable, and is indeed a substantial part of the truth about reference.

Be that as it may, a local descriptivism is disappointingly modest. It tells

[7] I owe the point to H. W. Noonan, 'Rigid Designation', *Analysis* 39 (1979). pp. 174-182.

[8] Hartry Field, 'Theory Change and the Indeterminacy of Reference', *Journal of Philosophy* 70 (1973), pp. 462-481.

us how to get more reference if we have some already. But where did the old language get *its* reference?

It is therefore tempting to try the same method on a grander scale. We can introduce terms in little families. How about bigger families? How about the biggest family of all—the entire vocabulary of the language? Then we needn't worry how the old vocabulary got its reference. Because there isn't any old vocabulary. (Or perhaps the old vocabulary is just the first-order logical vocabulary. Putnam seems to assume this, but without telling us why that vocabulary is special, or how it got its reference.) We go on just as before. The intended interpretation will be the one, if such there be, that makes the term-introducing theory come true. (Or: . . . come near enough to true. Or: the intended interpretation*s* will be the one*s*, 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.

And it leads straight to Putnam's incredible thesis. For *any* world (almost), whatever it is like, can satisfy *any* theory (almost), whatever it says. We said: 'the intended interpretation will be the one, *if such there be*, ...' Never mind the proviso—there *will* be. It is (almost) certain that the world will afford the makings of an interpretation that will make the theory come true. In fact, it will afford countless such interpretations. *Ex hypothesi* these interpretations are intended. So there is (almost) no way that the theory can fail to come true on its intended interpretations. Which is to say: (almost) no way that the theory can fail to come true *simpliciter*. This is Putnam's so-called 'model-theoretic argument'.[9]

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. It may be part of the truth about reference, but it cannot be the whole story. There must be some additional constraint on reference: some constraint that might, if we are unlucky in our theorising, eliminate *all* the allegedly intended interpretations that make the theory come true.

Further Constraints—Just More Theory? Putnam has constraints to offer: he speaks often of 'operational and theoretical constraints' (for instance, see R&R, p. 126; M&R, pp. 466, 469, 471, and 473). It is hard to tell from his words whether these are supposed to constrain reference or theory. Probably he thinks they do both: they constrain ideal theory, ideal theory is the term-introducing descriptive theory to which global descriptivism applies, so in this indirect way they constrain reference also. So these constraints work within global descriptivism. They are not an addition or alternative to it.

[9] The argument was anticipated (apart from mathematical detail having to do with the qualification '*almost* any world') in M. H. A. Newman, "Mr. Russell's 'Causal Theory of Perception'", *Mind* 37 (1928), pp. 137-148. Newman's argument is discussed in William Demopoulos and Michael Friedman, 'The Concept of Structure in Early Twentieth Century Philosophy of Science', in *Minnesota Studies in the Philosophy of Science* (forthcoming).

We must seek elsewhere for salvation from indeterminacy and over-easy truth. We need further constraints.

Putnam thinks there can be no such further constraints. Global descriptivism is the only possible account of reference (apart from accounts that rely on supernatural aid). Constraints that work within it are the only possible constraints on reference. His reason is that global descriptivism is imperialistic: it will annex any satisfactory alternative account of constraints on reference.

Suppose that we say it is constraint *C* that saves the day—a causal constraint, perhaps, or what have you. We offer an account of how constraint *C* works, a bit of theory in fact. If this bit of theory looks good, it will deserve to be incorporated into total theory. Suppose it is. Then an intended interpretation must make *C*-theory come true, along with the rest of total theory. But it will still be true, as much as ever, that (almost) any world can satisfy (almost) any theory. Adding *C*-theory to the rest of total theory doesn't help. It is still trivially easy for a world to make total theory come true, and in fact to do so in countless ways. And the point is general: it applies to any constraint (or, at least, to any otherwise satisfactory constraint) that might be proposed. Constraint *C* is to be imposed by accepting *C*-theory, according to Putnam. But *C*-theory is just more theory, more grist for the mill; and more theory will go the way of all theory.

To which I reply: *C* is *not* to be imposed just by accepting *C*-theory. That is a misunderstanding of what *C* is. The constraint is *not* that an intended interpretation must somehow make our account of *C* come true. The constraint is that an intended interpretation must conform to *C* itself.

That is why global descriptivism does not automatically annex its successful rivals. That is why global descriptivism, unaided by further constraints, is not the only possible theory of reference. That is why some further constraint on reference might save the day. Since Putnam's paradoxical thesis is patently false, we can be confident that there is indeed some further constraint, whether or not we can find out what it is.

Is that all? What I have just said (and others before me, *e.g.* Devitt,[10] in the course of advocating particular constraints) may not carry conviction. It may seem that Putnam is onto something deep and right. He is not just missing an easy distinction: satisfying *C*-theory versus conforming to *C*. Is there really a distinction here?

I think there is. But there are two reasons for doubting the distinction. One is simply misguided; the other is instructively wrong.

The misguided reason comes from the dialectic of philosophy. The rules of disputation sometimes give the wrong side a winning strategy. In particular, they favour the sceptic. They favour the ordinary sceptic about empirical knowledge; they favour the logical sceptic, Carroll's tortoise or a present-day doubter of non-contradiction; and they favour the sceptic about determinate reference. It goes as follows. The Challenger asks how determinate

[10] Michael Devitt, 'Realism and the Renegade Putnam: A Critical Study of *Meaning and the Moral Sciences*', *Noûs* 17 (1983), pp. 291-301.

reference is possible. The Respondent answers by giving an account of his favourite constraint. The Challenger says: 'Unless the words of your answer had determinate reference, you have not answered me unequivocally. So I challenge you now to show how the words of your answer had determinate reference. If you cannot, I can only take you to have proposed an addition to total theory—*that* I can understand, but that is futile.' If the Respondent answers just as before, he begs the question and loses. If he answers differently, he does not win, for he gets another challenge just like the one before. And so it goes. The Challenger is playing by the rules, and the Respondent cannot win. And yet the Respondent may indeed have given a correct account of the constraint that makes determinate reference possible, couched in language that does indeed have determinate reference in virtue of the very constraint that it describes! (Here I follow Devitt (*op. cit.*), generalising his account of the dialectical deadlock in case a causal constraint is proposed.) Moral: truth is one thing, winning disputations is another.

But there is a deeper and better reason to say that any proposed constraint is just more theory. Take your favourite theory of reference. Let us grant that it is true. But let us ask: what makes it true? And the tempting answer is: *we* make it true, by our referential intentions. We can refer however we like—language is a creature of human convention—and we have seen fit to establish a language in which reference works *thus*. Somehow, implicitly or explicitly, individually or collectively, we have made this theory of reference true by stipulation. '*We* interpret our languages or nothing does' (M&R, p. 482).

The main lesson of Putnam's Paradox, I take it, is that this purely voluntaristic view of reference leads to disaster. If it were right, any proposed constraint *would* be just more theory. Because the stipulation that establishes the constraint would be something we say or think, something we thereby add to total theory.

Referring isn't just something we do. What we say and think not only doesn't settle what we refer to; it doesn't even settle the prior question of *how* it is to be settled what we refer to. Meanings—as the saying goes—just ain't in the head.

What Might the Saving Constraint Be? Many philosophers would suggest at once that the saving constraint has to do with the causal chains that lead into the referrer's head from the external things that he refers to. At a minimum, some interpretations would be disqualified on causal grounds, and global descriptivism would select from those remaining. Or perhaps a causal account of reference ought to overthrow global descriptivism altogether.

If we subject a causal theory of reference (or a more modest causal constraint) to the 'just more theory' treatment, we get what I call causal descriptivism. That is: descriptivism, global or local, in which the descriptions are largely couched in causal terms. The lesson of Putnam's Paradox for causal theorists of reference is: don't trade in your genuine causal theory for causal descriptivism. But I myself would prefer causal descriptivism over a genuine causal theory. The causal theory often works, but not as invariably as philosophers nowadays tend to think. Sometimes an old-fashioned descriptivism works better; sometimes there are puzzling intermediate cases in which causal and descriptive considerations seem to tug in opposite directions.[11] When causal theories work, causal descriptivism works too.[12] When not, we need mixed theories, halfway houses between the 'new theory of reference' and the old. Causal descriptions seem ideally suited to mix into clusters with noncausal descriptions.

Given my preference for causal descriptivism—which indeed is just more description, just more theory—I must seek elsewhere for my saving constraint. I am inclined to favour a different kind of constraint proposed by G. H. Merrill.[13] (More precisely, he advises realists to propose it, but notes that he himself is no realist.) This constraint looks not to the speech and thought of those who refer, and not to their causal connections to the world, but rather to the referents themselves. Among all the countless things and classes that there are, most are miscellaneous, gerrymandered, ill-demarcated. Only an elite minority are carved at the joints, so that their boundaries are established by objective sameness and difference in nature. Only these elite things and classes are eligible to serve as referents. The world—any world—has the makings of many interpretations that satisfy many theories; but most of these interpretations are disqualified because they employ ineligible referents. When we limit ourselves to the eligible interpretations, the ones that respect the objective joints in nature, there is no longer any guarantee that (almost) any world can satisfy (almost) any theory. It becomes once again a worthy goal to discover a theory that will come true on an eligible interpretation, and it becomes a daring and risky hope that we are well on the way toward accomplishing this.

Merrill makes eligibility an all-or-nothing matter; I would prefer to make it a matter of degree. The mereological sum of the coffee in my cup, the ink in this sentence, a nearby sparrow, and my left shoe is a miscellaneous mess of an object, yet its boundaries are by no means unrelated to the joints in nature. It is an eligible referent, but less eligible than some others. (I have just referred to it.) Likewise the metal things are less of an elite, eligible class than the silver things, and the green things are worse, and the grue things are worse still—but all these classes belong to the elite compared to the countless utterly miscellaneous classes of thing that there are. *Ceteris paribus,* an eligible interpretation is one that maximises the eligibility of referents overall. Yet it may assign some fairly poor referents if there is good reason to. After all, 'grue' is a word of our language! *Ceteris* aren't *paribus,* of course; overall

[11] Such cases are presented in Peter Unger, 'The Causal Theory of Reference', *Philosophical Studies* 43 (1983), pp. 1-45.

[12] Even Saul Kripke grudgingly admits this: see footnote 38 to 'Naming and Necessity' in D. Davidson and G. Harman, eds., *Semantics of Natural Language* (Reidel, 1972). However, he doubts that a non-circular theory of *either* sort exists.

[13] G. H. Merrill, 'The Model-Theoretic Argument Against Realism', *Philosophy of Science* 47 (1980), pp, 69-81. For further discussion of Merrill's solution, see the final section of my 'New Work for a Theory of Universals'.

eligibility of referents is a matter of degree, making total theory come true is a matter of degree, the two desiderata trade off. The correct, 'intended' interpretations are the ones that strike the best balance. The terms of trade are vague; that will make for moderate indeterminacy of reference; but the sensible realist won't demand perfect determinacy.[14]

There seems to be a problem. To a physicalist like myself, the most plausible inegalitarianism seems to be one that give a special elite status to the 'fundamental physical properties': mass, charge, quark colour and flavour, . . . (It is up to physics to discover these properties, and name them; physicalists will think that present-day physics at least comes close to providing a correct and complete list.) But these elite properties don't seem to be the ones we want. Only in recent times have we had words for quark colour and flavour, but we have long had words for sticks and stones, cats, books, stars, . . . The solution, I suggest, is that we used to lack words for some very eligible referents because the correct interpretations of our language were the ones that did best on balance, not the ones that did best at best. Indeed, physics discovers which things and classes are the most elite of all; but others are elite also, though to a lesser degree. The less elite are so because they are connected to the most elite by chains of definability. Long chains, by the time we reach the moderately elite classes of cats and pencils and puddles; but the chains required to reach the utterly ineligible would be far longer still.

It is not to be said that our theorising makes the joints at which the world is to be carved. That way lies the 'just more theory' trap. Putnam would say: "very well, formulate your theory of 'objective joints in nature', what they are and where they are; and stipulate if you will that your referents are to be 'eligible'. But total theory with this addition goes the way of all theory: it is satisfiable with the greatest of ease in countless ways. And these countless ways, of course, assign countless different extensions to 'joint in nature', 'eligible', and the rest." No: the proposed constraint is that referents are to be eligible, not just that eligibility-theory is to be satisfied somehow, not just that the referents of 'cat' etc. are to be included among the referents of 'eligible'.

If I am looking in the right place for a saving constraint, then realism needs realism. That is: the realism that recognises a nontrivial enterprise of discovering the truth about the world needs the traditional realism that recognises objective sameness and difference, joints in the world, discriminatory classifications not of our own making. I do not quite say that we need traditional realism about universals.[15] For perhaps a nominalism that takes objective resemblance as primitive could do the job instead. But we need something of that sort. What it takes to solve Putnam's paradox is an objective inegalitarianism of classifications, in which grue things (or worse) are not

[14] It is not clear how much indeterminacy might be expected to remain. For instance, what of Quine's famous example? His rabbit-stages, undetached rabbit-parts, and rabbit-fusion seem only a little, if any, less eligible than rabbits themselves.

[15] As it might be, the theory of D. M. Armstrong, *Universals and Scientific Realism* (Cambridge University Press, 1978), discussed in my 'New Work for a Theory of Universals'.

all of a kind in the same way that bosons, or spheres, or bits of gold, of books are all of a kind.

I take it that Putnam classes the solution I advocate with solutions that rely on supernatural graspings or intuitings. He assimilates the view that 'the world . . . sorts things into kinds' to the preposterous view that the world gives things their names (RT&H, p. 53)! Recently, he has called my talk of elite classes 'spooky' and 'medieval-sounding'.[16] Well, sticks and stones may break my bones. . . . Anyway, what's wrong with sounding medieval? If the medievals recognised objective joints in the world—as I take it they did, realists and nominalists alike—more power to them. But I don't suppose that inegalitarianism of classifications is an especially medieval notion—rather, egalitarianism is a peculiarity of our own century.

Putnam has also said that inegalitarianism of classifications is contrary to physicalism. That would bother me, if true. But what's true is the opposite: *egalitarianism* is contrary to physicalism. For physicalists take physics—as it now is, or as it will be—at face value. And physics professes to discover the elite properties. What is the content of this part of physical science, according to an egalitarian?[17]

*   *   *

That completes the first part of the paper. Now I shall take up five questions about why Putnam proceeds as he does—questions that leave me uncertain how well I have understood what he is up to.

Why 'Model-Theoretic'? The premise that joins with global descriptivism to yield disaster is *not* any big theorem of model theory. In particular, it is not the theorem that gets star billing in M&R, the Skolem-Löwenheim Theorem. In fact, what's needed is pretty trivial. As I put it before: (almost) any world can satisfy (almost) any theory. The first 'almost' means 'unless the world has too few things'; the second means 'unless the theory is inconsistent'. This premise is obtained as follows. A consistent theory is, by definition, one satisfied by some model; an isomorphic image of a model satisfies the same theories as the original model; to provide the makings for an isomorphic image of any given model, a domain need only be large enough.

The real model theory adds only a couple of footnotes that are not really crucial to the argument. First, by the Completeness Theorem, we could if we wished redefine 'consistent' in syntactic terms. Second, by the Skolem-Löwenheim Theorem, our 'unless the world has too few things' is less of a qualification than might have been supposed: any infinite size is big enough. But the qualification wasn't very important in the first place. If Putnam's thesis had been that an ideal theory can misdescribe the world only by getting

[16] In remarks presented at the annual conference of the American Philosophical Association, Eastern Division, Baltimore, 1982.

[17] See the discussion of formulations of materialism in my 'New Work for a Theory of Univerals', in which I argue that inegalitarianism of classifications must be presupposed in stating materialism.

its size wrong, that would have been incredible enough. And in fact that *is* Putnam's thesis: for all he has said, it is still possible for ideal theory to misdescribe a finite world as infinite. Who cares whether the possibility of similar mistakes among the infinite sizes also is granted? We thought it was possible to misdescribe the world in ways having nothing to do with its size.

Anyway, the applicability of the model theory depends on treating exactly the first-order logical vocabulary as 'old' language, with antecedently determinate reference. As Robert Farrell[18] has emphasised, Putnam has no right to give this vocabulary special treatment. Perhaps he only did it for the sake of the argument, giving away points just because he would not need them.

Why Just Ideal Theories? You should have spotted a shift in my formulations of Putnam's incredible thesis. The official formulation was this: it makes no sense to suppose that an *empirically ideal* theory might nevertheless be false. But the conclusion of the model-theoretic argument applies to *any* consistent term-introducing total theory to which global descriptivism applies. It makes no sense (small worlds aside) to suppose that *any* such theory might be false, whether or not it is ideal. Idealness of the theory doesn't figure in the proof.

Perhaps Putnam has chosen to underplay his hand. Perhaps he does think of the model-theoretic argument as showing that our total accepted theory cannot be false, whether or not it is ideal (unless it is inconsistent, or the world is too small). But the special case of an ideal theory is the case that distinguishes realists from Peirceans, so that is the case he chooses to discuss. This hypothesis nicely fits the text of R&R, pp. 125-126. Even so, I think it is most likely a misunderstanding.

For one thing, why does he pass up the opportunity to say that the 'incoherent picture' held by his realist opponents commits them to the absurdity that even non-ideal theories are true on their intended interpretations, if that is what he thinks?

More likely, the model-theoretic argument is supposed to work only for ideal theories. But how could that be? A theory does *not* need to be ideal, merely consistent, in order to be satisfiable in any (big enough) world. So maybe the first premise of the argument, global descriptivism, is only supposed to work for ideal theories. But then how could it say anything about the vocabulary of our actual, present total theory, which doubtless isn't quite ideal? Perhaps as follows.

Descriptivism, local or global, might be *futuristic.* That is, the term-introducing theory which is supposed to come true on intended interpretations, if such there be, might be not the theory by which the terms actually were introduced, but rather some improved descendant that is expected to exist in the future. It might even be some ideally improved descendant that is never expected to actually exist, but that would result if the process of improvement went on forever. Imagine that our detective says: "My present

[18] Robert Farrell, 'Blanket Skolemism', presented at the annual conference of the Australasian Association of Philosophy, Sydney, 1980.

hypothesis is that Jack the Ripper—as I propose to call him—did this, that, and the other. Of course, I realise that most likely that's not quite right. But if we start with this hypothesis, and improve it bit by bit in accordance with the evidence and the canons of scientific detection, eventually we may have a Jack-the-Ripper theory that can be improved no further. Maybe we will have it; maybe we never will, but I can speak of it even so. By 'Jack the Ripper' I intend to refer to the one described by that ultimate Jack-the-Ripper theory that we may never see." Likewise there could be a futuristic global descriptivism. Perhaps that is what Putnam has in mind. (I don't see any explicit futurism in R&R; I do in RT&H, pp. 30-32, but not in the context of the model-theoretic argument; in M&R, p. 475, there is explicit consideration of futuristic and non-futuristic alternatives.) I think it is what he should have in mind, for a reason to be stated shortly.

Why Anti-Realist? Why does the model-theoretic argument attack realism? By definition, of course: 'It is this feature [that an ideal theory might be false] that distinguishes metaphysical realism, as I am using the term, from the mere belief that there *is* an ideal theory . . .' (R&R, p. 125). But what makes *that* a definition of any form of *realism*?

My point is emphatically not that 'Internal realism is all the realism we want or need' (R&R, p. 130). Internal realism, I take it, is realism feigned. The plan is to speak exactly as the realists do (except in the philosophy room—I have no idea how that lapse can be justified); and to do so in good conscience, in the hope that one's words are destined to join the ideal theory, and so are 'epistemically true'; but to do so without any intention of describing the world by saying something that will be true only if the world is one way rather than the other. (But of course the Internalist will *say* that he intends to be 'describing the world . . .'. His plan is to speak *exactly* as the realists do!)

My point is rather that even if the model-theoretic argument worked, it would not blow away the whole of the realist's picture of the world and its relation to theory. Something vital would be destroyed, but a lot would be left standing. There would still be a world, and it would not be a figment of our imagination. It would still have many parts, and these parts would fall into classes and relations—too many for comfort, perhaps, but too many is scarcely the same as none. There would still be interpretations, assignments of reference, intended and otherwise. Truth of a theory on a given interpretation would still make sense, and in a non-epistemic way. Truth on all intended interpretations would still make sense. Despite Putnam's talk of the 'collapse' of an 'incoherent picture', he has given us no reason to reject any of these parts of the picture. The only trouble he offers is that there are too many intended interpretations, so that truth on the intended interpretations is too easily achieved. That is trouble, sure enough. But is it *anti-realist* trouble, except by tendentious definition? It seems to me exactly opposite to traditional anti-realism.

The traditional anti-realist doubts or denies that there is any world save a figment of our imagination. Or he doubts or denies that the world divides into parts except insofar as we divide it, or that those parts fall into any classes or relations except such as are somehow of our own making. Or he doubts or denies that we can achieve reference to parts of the world, he questions that there can be even one intended referential interpretation. Or he doubts or denies that we can ever achieve truth on intended interpretations, or that we can ever have reason to believe that we have done so.

Across the board, wherever traditional anti-realism sees privation, Putnam argues instead from overabundance. It is only at the end that the opposites meet. They agree that it is unreasonable for science to aim at accurate description of reality, as opposed to the 'epistemic truth' of ideal theory. But why is that? Is it because accurate description is so difficult that we could not attain it, or could not reasonably expect to, even if we attain 'epistemic truth'? Or is it rather because accurate description is easy, automatically attained along with 'epistemic truth' and adding nothing extra?

Putnam should say the latter. He gives us no argument that discredits the realist's conception of truth of a theory on an interpretation which assigns referents in the world. His strategy should be to co-opt that conception, not to oppose it. He ought to say: '*Contra* realist orthodoxy, truth *simpliciter* is equivalent to, or simply is, epistemic truth. That is not because there is anything epistemic about truth-on-an-interpretation. Nor is it because truth *simpliciter* is anything else than truth on all intended interpretations. Rather it is because intendedness of interpretations is an epistemic matter.'

Maybe this *is* Putnam's strategy. His presentations of the model-theoretic argument in R&R and M&R can very well be read accordingly. If so, then global descriptivism needs to be futuristic, else truth on all intended interpretations will coincide with 'epistemic truth' only in the sweet by and by when we have ideal theory. This is my postponed reason why I think Putnam should have the futuristic version of global descriptivism in mind.

But if Putnam's strategy really is as I have just imagined, then there is a lot of poetic licence in some of what he says. It is an exaggeration to say that the realist picture 'collapses' (R&R, pp. 126 and 130). That suggests destruction more total than has actually been accomplished. And it is quite uncalled for to say, however metaphorically, that 'the mind and the world jointly make up the mind and the world' (RT&H, p. xi). No; we make theories, not worlds. The *metaphysics* of realism survives unscathed. What does suffer, if Putnam has his way, is realist semantics and epistemology.[19]

Why are Supernatural Constraints Exempt? Putnam presents the model-theoretic argument as bad news for moderate, naturalistic realists: 'it is, unfortunately, the *moderate* realist position which is put into deep trouble . . .' (M&R, p. 464). Verificationists who aspire only to 'epistemic truth' have nothing to fear. But neither, he says, do those immoderate realists who claim to achieve determinate reference by supernatural means—by grasping, by intuiting, by direct contact, by magic, by noetic rays, by sixth sense, call it

[19] In this section I am indebted to Devitt's insistence that it is really very peculiar to take realism as an issue about semantic theory.

what you will. *Their* only problem is that their views are scientifically disreputable.

Why is that? Is it just that Putnam magnanimously declines to fight the weak? Or would supernatural intercourse between thinker and referent actually afford some way around Putnam's argument? I do not see how supernatural acquaintance with referents could do any better than the natural sort. Why is it a better way to achieve determinate reference if we get cat Nana into the grasp of our noetic rays than if we hold her in our hands? Why is it better if we intuit her with our sixth sense than if we see her with our eyes and hear her with our ears?

We know what Putnam says if we try to base determinate reference on natural causal connection: the theory of the causal constraint on reference is just more theory, as subject as any theory to overabundant, conflicting intended interpretations. But why are supernatural constraints exempt from parellel treatment? What's the good of holding up yet another sign, thus

$$\boxed{\text{DIRECTLY GRASPS}}$$

or perhaps

$$\boxed{\text{INTUITS}}$$

if it is still open to Putnam to challenge the determinate reference of the words written on the sign? (*Cf.* R&R, p. 127) What can the proposed supernatural constraint be, if not the useless requirement that grasping-theory, or whatever, shall be made to come true along with the rest of (futuristic?) total theory?

I have argued, of course, that it is fair to reject the 'just more theory' treatment, whatever constraint it may be applied to. Presumably Putnam disagrees. But he has said nothing to show why the treatment applies only to natural constraints.

Perhaps Putnam thinks that supernaturalists are immune from the 'just more theory' treatment because they deny the premise that '*we* interpret our languages or nothing does', or in other words that constraints on reference obtain only because we stipulate that they do. That would be a good reason to grant them exemption. I reply that a naturalist also can deny it on behalf of natural constraints, as I have done.

What is the Vat Argument? In R&R, setting forth the picture to be refuted by the model-theoretic argument, Putnam mentions brains in a vat:

> . . . indeed, it is held [by the metaphysical realist] that we might be *unable* to represent THE WORLD correctly at all (*e.g.* we might all be 'brains in a vat', the metaphysical realist tells us). (p. 125)

And a little later, still in connection with the model-theoretic argument,

> Suppose we . . . are and always were 'brains in a vat'. Then how does it come about that *our* word 'vat' refers to *noumenal* vats and not to vats in the image. (p. 127)

So when RT&H opens with a discussion of brains in a vat, we know what to expect. Brains in a vat are a stock example of radical deception. The model-theoretic argument is meant to show that radical deception is impossible. (More precisely: that radical deception is possible only when the deceived fall radically short of 'epistemic truth', which *ex hypothesi* the brains do not.) We expect Putnam to introduce the model-theoretic argument in dramatic fashion by using it to argue that even a brain in a vat is not radically deceived; and hence that it is nonsense to fear that *we* are radically deceived brains in a vat.

We might also expect that Putnam would anticipate the objection that the model-theoretic argument ignores causal constraints on reference; and that he might wish to postpone his 'just more theory' rejoinder, since it is tricky and he is obviously writing in part for nonspecialists; and that he might therefore offer an interim reply to the advocate of causal constraints, a reply that works only in this special case. And he does have such a reply. Even if a causal theory is correct and relevant—*contra* the 'just more theory' rejoinder—it doesn't help in this case. For it tends to show that we and the brains do not refer to the same things when we use the same words, since they are causally isolated from our referents. (Here I imagine Putnam to concede temporarily that our own reference is governed at least partly by causal constraints.) So at least some apparent examples of deception are mitigated, rather than worsened, by applying causal constraints on reference. If the brain says (in his inner speech) 'I am in Vienna', we might carelessly suppose that he means what we do and is therefore mistaken. (The brain and his vat are not in Vienna.) For he cannot think of Vienna, for he cannot refer to it, for *ex hypothesi* he is causally isolated from it.

All this, I repeat, is what we might reasonably expect. It is not what Putnam gives us. The argument in Chapter 1 of RT&H is not the model-theoretic argument.[20] The causal theory of reference is not used *ad hominem* against a hypothetical objector who hopes to use it to defend determinate reference. It is defended vigourously, then it carries the whole weight of the argument that the brains in the vat are not deceived.

Putnam's defence of the causal theory of reference is fair. Even if he really thinks it is just more theory, and couched in language with radically indeterminate reference, he may still think it is good theory, in all probability 'epistemically true'. Surely he does think so.

But how can the causal theory of reference, unaided, carry the whole weight of the argument? I see how it can be used to exonerate the brains from various specific accusations of error. "The brain says that by 'Vienna' he refers to Vienna and he doesn't."—"No, he *does* refer to what *he* calls 'Vienna', and

[20] For a genuine model-theoretic argument that brains in a vat are not deceived, see Paul Horwich, 'How to Choose Between Empirically Equivalent Theories', *Journal of Philosophy* 74 (1982), pp. 61-77.

so he speaks the truth. For what he calls 'Vienna' isn't the city that he is isolated from, but rather is part of the computer program that is the source of his 'Vienna'-tokens and his mental 'Vienna'-dossier."

Or even, perhaps, "He says that he isn't a brain in a vat, but he is." – "No, he's right; for by 'vat' he means not the sort of thing he is in, but rather the sort of thing that is the source of this 'vat'-tokens and his 'vat'-dossier. That sort of thing is, again, a sort of part of the computer program, and he isn't in one of those."

So far, so good. But it's not good enough just to show that the brain doesn't make certain specific errors – he might make ever so many other errors, and be very radically deceived indeed.

In fact, showing that the brain avoids certain specific errors *is* enough to meet Putnam's stated goals in Chapter 1 of RT&H. He doesn't promise more than he delivers. But it isn't enough to contribute to Putnam's overall strategy. So what if we have to be a bit careful in saying just what mistakes a radically deceived brain in a vat does and doesn't make? No worries for realism there! Anybody can grant the point. I willingly do.

I suspect that Putnam thinks that the causal theory of reference can be used over and over, in similar fashion to the examples just considered, to exonerate the brain from all accusations of error whatsoever. (Except when the brain falls short of 'epistemic truth'.) If that is what Putnam means us to think, then Chapter 1 of RT&H fits nicely into his plan of battle. The causal-theoretic argument against massive error takes its place as partner to the model-theoretic argument, at least for a certain class of cases.

But I think we have been given no reason whatever to suppose that causal-theoretic exonerations can be produced *en masse*. The requisite causal theory of reference does not exist. We have a fairly well developed causal theory of reference for ordinary proper names, *e.g.* of people and places;[21] and we have a sketchy causal theory of reference for names of chemical substances and plant and animal species. The causal theory of reference for the whole of language is *not* just over the next hill. We don't know whether there is any such theory to be had, still less how credible it would seem or what it would say.

Suppose I accuse the brain of error when he says 'I am in Vienna', not because he isn't in the city that *we* call 'Vienna' but because he isn't *in* that part of a computer program that *he* calls 'Vienna'. You appeal to the causal theory of reference for the word 'in', saying that he *is* in the relation that *he* calls 'in' to the appropriate piece of program. I say it's a fraud. You may hope someday to possess a credible causal theory of reference that yields that result, but I do not believe you now possess even a glimmering of how this theory will go.

(One reason to doubt that you will ever have it: whatever happens in special cases, causal theories usually make it *easy* to be wrong about the things we refer to.)

[21] I have in mind the theory of Michael Devitt, *Designation* (Columbia University Press, 1981). But I don't endorse even Devitt's theory; recall that I'm inclined to prefer causal descriptivism.

And so it will go, I think, time and again. The exonerations of error to be had from extant causal theories of reference will be few and far between. The model-theoretic argument against massive error is fatally flawed, I think; but the causal-theoretic argument is barely a starter.

Princeton University

Received June 1983