Putnam, Koethe, and Metaphysical Realism

Shekhar Pradhan
University of Illinois at Urbana-Champaign

I

In a discussion note titled "Putnam's Argument Against Realism"[1] John Koethe attempts to refute Putnam's main argument against a view which Putnam calls 'Metaphysical Realism'. (This argument was first put forward by Putnam in Realism and Reason",[2] and in the philosophical literature has come to be called the model-theoretic or modelling argument against Metaphysical Realism). While I am not at all convinced by Putnam's argument, I find Koethe's refutation of the argument even less convincing. In this note I attempt to show where Koethe has gone wrong.

According to Putnam, Metaphysical Realism (hereafter abbreviated as M.R.)is more a picture or a model than a theory. It purports to be "a model of the relation of any correct theory to all or part of THE WORLD."[3] According to this model, "there has to be a determinate relation of reference between terms in L and pieces (or sets of pieces) of THE WORLD."[4] The most important consequence of M. R. is that truth becomes radically non-epistemic; that is to say, according to M. R., even a theory that is ideal in every way might turn out to be false. This consequence of M. R. Putnam finds to be unintelligible, and the modelling argument is designed to show just that, viz. that this consequence of M. R. lacks intelligibility.

Let T_1 be an ideal theory by our lights. Now, the argument against the intelligibility of the claim that T_1 might be false runs thus:

> I assume 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 (by hypothesis) and has (only) infinite models. So by the completeness theorem (in its model theoretic form), 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 L and sets of pieces of THE WORLD--such that the theory T_1 comes out <u>true</u>--true of THE WORLD--provided we just interpret 'true' as TRUE (SAT).[5]

Furthermore, according to Putnam, there are no grounds for ruling out SAT as an unintended correspondence. And, therefore, "the supposition that even an 'ideal' theory (from a pragmatic point of view) might really be false appears to collapse into unintelligibility."[6]

II

It is this argument that Koethe attempts to refute. Koethe's attack is focused on the notion of an ideal theory invoked in the above argument. His contention is that either there is no coherent conception of an ideal theory which suits Putnam's argument or it is a sense of ideal theory such that in this sense of an ideal theory the realist is <u>not</u> obliged to hold that even an ideal theory might be false. Crucial to the argument for this disjunction is Koethe's claim that in order for the model-theoretic argument to work the theory T_1 has to be one that is unrevisable-in-principle. And Koethe goes on to argue that if this last claim is true, then an important ingredient in the notion of ideal theory that Putnam requires for his argument has to be unrevisability. According to Koethe, it is this necessary ingredient in that notion of ideal theory that Putnam needs which makes certain ways of understanding the notion of an ideal theory, which may be otherwise unobjectionable, inadequate for Putnam's purposes. On the other hand, according to Koethe, if one constructs a notion of an ideal theory that has this ingredient and is otherwise unobjectionable then it is not clear that on this way of explicating (or constructing) the notion of an ideal theory the Metaphysical Realist is committed to holding that the ideal theory might be false.

My claim is that Koethe has gone wrong in holding that in order for the model-theoretic argument to work the theory T_1 has to be one that is unrevisable-in-principle, and as a consequence Koethe has unfairly dismissed certain conceptions of ideal theory as inadequate for Putnam's purposes.

Let us consider Koethe's argument for the claim that any theory for which the model-theoretic argument works has to be unrevisable-in-principle. His strategy is "first try to apply the argument not to an ideal theory, but to an arbitrary member of a sequence of theories each of which is, in the course of time, accepted and then abandoned in favor of a successor."[7]

Let T_t be an arbitrary member of such a sequence of theories. Now let us see if we can render unintelligible the claim that T_t might be false by applying the model-theoretic argument to it. Such an application would run as follows:

> Pick a model M (we would argue) of the same cardinality as THE WORLD. Map M's domain one-to-one into pieces of THE WORLD, and use this mapping to define the relation of M directly in THE WORLD. We now have a relation SAT between the terms of the language in which T_t is expressed and THE WORLD, on the basis of which we can define a property TRUE (SAT) such that T_t comes out true only if we interpret 'true' as TRUE (SAT). And (we would claim) there are no legitimate grounds for rejecting this interpretation of 'true' as unintended.[8]

However, according to Koethe, there <u>is</u> a reply to this argument. He claims that <u>there are</u> grounds for rejecting TRUE (SAT) as an acceptable interpretation of 'true' for T_t. There are two lines of argument that Koethe gives for this claim between which he does not distinguish, and he may well be confusing these two lines.[9] They are:

Argument A): If holders of T_t accept TRUE (SAT) as the intended interpretation of 'true' for T_t, they arrive at an absurd consequence, and, therefore, they will have legitimate grounds for rejecting it as the intended interpretation.[10]

And

Argument B): Holders of T_t are in a position to know at t that later theorists at t+1, if there are any, will have to reject TRUE (SAT) as the intended interpretation for their theory (that is, T_{t+1}) and that gives the holders of T_t legitimate basis for ruling out TRUE (SAT) as the intended interpretation of 'true' for T_t.[11]

It is worth pointing out in passing that Koethe has concerned himself entirely with showing that the holders of T_t will have legitimate reasons for ruling out TRUE (SAT) as unintended rather than trying to show that it is <u>we</u>, the constructors of this thought experiment, the philosopher-theorists, who will have legitimate reasons for ruling out TRUE (SAT) as unintended. But Putnam's argument has to do with the latter claim rather than the former. So if Koethe's argument is not to be a complete <u>non sequitur</u> some such

principle as the following must be true: We can decide whether it makes sense to suppose that a theory T might be false (or decide that a certain interpretation of 'true' for T is unintended) by deciding whether the holders of T will have legitimate grounds for thinking it makes sense to suppose that T might be false (or, by deciding whether the holders of T will have legitimate grounds for rejecting that interpretation as unintended). However, this principle is not always true. Whether someone has legitimate grounds for thinking something depends on what they know and what they are ignorant of. So this principle will be false whenever we, the constructors of the thought-experiment have a relevant piece of information that the holders of the theory do not possess. What gives them legitimate grounds for holding something in their ignorance of this piece of information need not give us legitimate grounds for holding the same thing in our knowledge of this piece of information.

I shall briefly sketch and amplify the line of argument A and show that it is unsound, and then I shall go on to consider argument B in some detail.

Argument A

1) TRUE (SAT) is an acceptable interpretation of 'true' just in case SAT is an acceptable interpretation of 'reference' or 'satisfaction'.

1') Suppose SAT is an acceptable interpretation of 'reference' for T_t.

2) T_t will be replaced at a later time t+1 by T_{t+1}.

3) Referents of terms are for the most part preserved across change of theories. (Koethe borrows this premise from Putnam).

4) Therefore, if SAT is taken as the intended relation of reference between the terms of T_t and reality, then this must be the intended relation of reference between these same terms as occurring in T_{t+1} and reality (from 3).

5) Since T_{t+1} is inconsistent with T_t, T_{t+1} must be FALSE (SAT).

6) Holders of T_{t+1}, if there are any, will be in a position to know 1-5.

7) Therefore, if SAT is taken as the intended interpretation of 'reference' for T_t, then holders of T_{t+1} will have to accept SAT as the intended interpretation of 'reference' for T_{t+1}. And so they

will be holding a theory they know to be false. (From 4, 5 and 6).

8) Holders of T_t are ,at t, in a position to know 1-7.

9) Therefore, holders of T_t are in a position to realize that if SAT is taken as the intended interpretation of 'reference' for T_t, they have to allow as a consequence that later rational theorists, if there are any, will have to accept a theory they (the later theorists) know to be false. (From 7 and 8).

10) But this consequence is an absurdity.

11) Therefore, holders of T_t are in a position to know at t that if SAT is taken as the intended interpretation of 'reference', it leads to an absurdity. (From 9 and 10).

12) Therefore, this knowledge gives them (the holders of T_t) legitimate grounds for rejecting SAT as the intended interpretation of reference for T_t and for rejecting TRUE (SAT) as the intended interpretation of 'true' for T_t.

The most crucial part of this argument is the claim (number 7 above) that if SAT is taken as the intended interpretation of 'reference' for T_t, then holders of T_{t+1} will have to accept it for their theory too, even though they know that that makes T_{t+1} FALSE (SAT). But why should they have to do that? Being rational, they will reject SAT as the intended interpretation of 'reference' for T_{t+1}. And holders of T_t can figure out as much. So why does Koethe think that holders of T_t are committed to supposing that if SAT is taken by them as the intended interpretation of 'reference' for T_t, then holders of T_{t+1} have to hold SAT to be the intended interpretation of 'reference' for T_{t+1} too? (It is not clear Koethe holds this. But he does not clearly distinguish between this line of argument and line B, which will be described below.)[12]

The only answer that Koethe can give is that anyone who holds premise 3 and also holds that SAT is the intended interpretation of 'reference' for T_t is committed to holding that SAT is the intended interpretation of reference for T_{t+1} (since reference of most terms is supposed to be preserved in the transition from T_t to T_{t+1}). But 3 is open to two interpretations:

3') For most terms that the new theory (T_{t+1}) and the old theory (T_t) have in common, these terms as occurring in the new theory should (for the most part) be assigned the same referents as are assigned to these terms as occurring in the old

theory by an assignment of referents according to (or, in light of) the earlier theory (T_t).

And

3") For most terms that the new theory (T_{t+1}) and the old theory (T_t) have in common, these terms as occurring in the new theory should (for the most part) be assigned the same referents as are assigned to these terms as occurring in the old theory by an assignment of referents according to (or, in light of) the later theory (T_{t+1}).

Notice that 3' is false, whereas 3" may well be true. And more to the point, it is not 3' that Putnam holds, but 3".[13] So if Koethe is to use a picture of theory-evolution acceptable to Putnam, he must hold the second interpretation of 3. But also notice that for the above argument to go through Koethe requires 3' rather than 3". It is only those who hold 3' who are committed (on holding SAT to be the intended interpretation of 'reference' for T_t) to holding that SAT be the intended interpretation of 'reference' for T_{t+1} too. On the other hand, those who hold 3" will conclude that later theorists will probably reject SAT as the intended interpretation for T_{t+1}, and they will also conclude that (since later theorists hold 3") later theorist will reject SAT as the intended interpretation of 'reference' for T_t too. So the absurd consequence that is supposed to follow from acceptance (by holders of T_t) of SAT as the intended relation of 'reference' for T_t will not follow and, hence, 7 and 9 above will be false. So, in conclusion, if 3 is understood as 3', the argument above will be unacceptable to Putnam (and is probably unsound) and if 3 if understood as 3", the argument above will be invalid because in that case 7 does not follow from 4, 5 and 6 or any of the preceding propositions.

The other line of argument, which I have called Argument B, makes use of the second interpretation of 3. It runs as follows:

Argument B

1) TRUE (SAT) is an acceptable interpretation of 'true' for T_t just in case SAT is an acceptable interpretation of 'reference' or 'satisfaction' for T_t.

2) T_t will be replaced at a later time t+1 by T_{t+1}.

3") For most terms that the new theory (T_{t+1}) and the old theory (T_t) have in common, these terms as occurring in the new theory should (for the most part) be assigned the same referents as are as-

signed to those terms as occurring in the old theory by an assignment of referents according to (or, in light of) the later theory (T_{t+1}).

4) Since T_{t+1} is inconsistent with T_t, if T_t is TRUE (SAT) then T_{t+1} must be FALSE (SAT).

5) Holders of T_{t+1}, if there are any, are in a position to know 1-5.

6) Therefore, being rational creatures, they will reject SAT as the intended interpretation of T_{t+1}.

7) Therefore, they will reject SAT as the intended interpretation of T_t (From 3", 5 and 6).

8) Holders of T_t are ,at t, in a position to know 1-7. In particular, they know that holders of T_{t+1} (if there are any) will reject SAT as the intended interpretation of T_t.

9) This gives holders of T_t legitimate grounds at t for rejecting SAT as the intended interpretation of 'reference' for T_t and for rejecting TRUE (SAT) as the intended interpretation of 'true' for T_t.

Furthermore, Koethe holds that it is not necessary that holders of T_t know that T_t will be replaced by T_{t+1}, it is enough that for all they know it is possible that T_t will be replaced by T_{t+1}. So, according to him, the above argument should work against any theory T_t such that its adherents are not in a position to know that it won't be replaced. That is to say, according to Koethe, this refutation works against any application of the model-theoretic argument to a theory that is open to revision. Only an application of the model-theoretic argument to a theory which is not open to revision--is unrevisable, and unrevisable in principle--escapes this refutation, or so Koethe claims. Therefore, a property an ideal theory such as T_1 for which the model-theoretic argument is supposed to work has to be that T_1 is unrevisable. Thus Koethe says, "it is important to emphasize that this reply does not require that T_t actually have an incompatible successor which is operationally verified at t+1. It is enough that the proponents of T_t are not in a position to know, at t, that it will not, and hence have to admit the possibility that rational later speakers will accept a theory inconsistent with T_t."[14]

III

The suspicious part in the above argument (B) is 9. Why should the fact (leave alone the possibility) that someone else is later going to reject their (i.e. the-

holders-of-T_t's) interpretation of 'reference' be sufficient reason for holders of T_t to reject their own interpretation of 'reference'? Isn't this slavish deference to the views of others uncalled for? Yet that is what 9 claims, and that is what Koethe would present to us as the paradigm of rationality. Clearly, Koethe couldn't have meant this.

Let us suppose that for an interpretation to be intended is for it to have some property (or set of properties) F, and for an interpretation to be unintended is for it to lack F. Let us also suppose that later theorists are in a better position (than holders of T_t) to know whether SAT as an interpretation of 'reference' for T_t has F or not. Now, if holders of T_t know that such later theorists who are in a better position to know whether SAT has F or not will pronounce that SAT lacks F, then in that case this rejection of SAT by later theorists will provide holders of T_t legitimate grounds for rejecting SAT as the intended interpretation of 'reference' for T_t. But holders of T_t are not likely to be in a position to know any such thing. At most, they can only speculate and conjecture this--slim grounds for rejecting their own interpretation as unintended. And even if we waive this objection, the above case only shows that where revision of a theory by later better-informed theorists (better-informed among other things about which interpretations of 'reference' for earlier theories possess F and which do not) occurs, there the holders of previous theories (who know that such a revision will take place) may take this future revision to be legitimate grounds for rejecting their interpretation of 'reference' as unintended.

But what about cases where no such revision of the theory takes place? Will the holders of that theory then have legitimate grounds for rejecting their interpretation of 'reference' for that theory as unintended? Koethe thinks that they will so long as they do not know that the theory won't be revised. That is, so long as the holders of T_t have to admit the possibility that T_t may come to be revised, this bare possibility gives them legitimate grounds for rejecting their interpretation as unintended. But surely this is far too strong. Why should the possibility that some better-informed person is going to reject my interpretation as lacking F give me reason to think that my interpretation in fact lacks F? Surely, there is nothing irrational in holding some proposition p and also admitting that it is possible--barely possible--that someone wiser may someday hold not-p. (And is this possibility ever absent? So, then, is it never rational for us to hold anything!) It is entirely incomprehensible to me why Koethe should think otherwise. Yet Koethe does say:

> [the holders of T_t] have a legitimate basis on which to reject the interpretation of 'true' as TRUE (SAT)--namely, that later speakers of the language (should there be any), or they themselves at some later time, probably could not accept it.[15]

And he goes on immediately to emphasize that

> this reply does not require that T_t actually have an incompatible successor which is operationally verified at t+1. It is enough that the proponents of T_t are not in a position to know, at t, that it will not, and hence have to admit the possibility that rational later speakers will accept a theory inconsistent with T_t. They cannot admit this possibility and at the same time accept the interpretation of 'reference' as SAT . . . .[16] (my emphasis)

The situation so far is thus: Putnam has given an argument--the model theoretic argument--applied to an 'ideal' theory T_i which argument is designed to show that it is unintelligible to hold that that theory might be ideal and yet be false. Koethe applies this argument to a theory T_t which is an arbitrary member of a sequence of theories arranged chronologically and which is (therefore?) supposed to be less than Ideal (Why Koethe assumes that such a theory is less than ideal I do not know). Then he produces an argument (or a confusion of two distinct lines of argument)[17] to show that in this case there is a way of ruling out TRUE (SAT) as an unintended interpretation of 'true' for T_t and so the model-theoretic argument cannot be applied to T_t. He thinks that this argument of his succeeds (in refuting the model-theoretic argument) in every case of a theory where the holders of that theory have to admit the possibility that that theory may come to be replaced by another theory (inconsistent with it). And so he claims that the model-theoretic argument works, if it works at all, only when applied to a theory that is unrevisable from the point of view of those who hold it. Then he goes on to argue that there can be no such theory.

In my response to Koethe I have been concerned to show that he is wrong in maintaining that his argument (A or B above, or a confused mixture of the two) succeeds in showing that the model-theoretic argument fails in every case of a theory where the holders of that theory have to admit the possibility that that theory may come to be replaced by another theory (inconsistent with it). I claim that all that Koethe's argument has succeeded in showing is that the model-theoretic argument fails in every case of a theory where the holders of that theory know that beings wiser

than them will in fact replace that theory by another theory (inconsistent with the earlier theory). And so, I claim that Koethe is wrong in holding that if the model-theoretic argument is to work in the case of an ideal theory such as T_I, then T_I must be unrevisable from the point of view of those who hold it. It is not that I think Putnam's notion of an ideal theory is clear or unproblematic; however, I do think that Koethe unfairly saddles Putnam's notion of an ideal theory with unrevisability. And for this reason Koethe's attack misses its mark. Putnam's answer to Koethe should be that the ideal theory does not have to be unrevisable, it is enough that the holders of the ideal theory do not know that beings wiser than them will replace their theory by a theory inconsistent with the ideal theory. And we, the constructors of the thought-experiment, can guarantee this by stipulating that the ideal theory be one that will in fact never be revised and so, of course, it cannot be part of anyone's knowledge that the ideal theory will be revised. (Of course its not being revised may have nothing much to do with the ideal theory being ideal--its idealness could be on other grounds).

NOTES

[1]John Koethe, "Putnam's Argument Against Realism," The Philosophical Review, LXXXVIII, 1 (January, 1979), pp. 92-99.

[2]Hilary Putnam, "Realism and Reason," Proceedings and Addresses of the American Philosophical Association, Vol. 50, No. 6 (August, 1977), pp. 483-497. Also reprinted in Putnam, Meaning and the Moral Sciences (London, 1978). All references to Putnam are to the Meaning and the Moral Sciences, hereafter abbreviated as MMS.

[3]MMS, p. 123.

[4]Ibid., p. 125.

[5]Ibid., pp. 125-126.

[6]Ibid., p. 126.

[7]Koethe, pp. 93-94.

[8]Koethe, pp. 94-95.

[9]To be fair to Koethe it must be pointed out that some crucial sentences are missing, due to a printer's error, from the passage on p. 95 where he presents his argument. This makes it difficult to tell which of the two lines of argument sketched below Koethe means to be advancing.

[10]Evidence for this line of argument can be found in passages such as: "They [the holders of T_t] cannot admit this possibility [the possibility that later rational speakers will accept a theory inconsistent with T_t] and at the same time accept the interpretation of 'reference' as SAT, for to do so would be to allow that rational later speakers might accept as operationally verified a theory which they know to be false (since any later speakers can be assumed to know that on this interpretation any theory inconsistent with T_t is FALSE (SAT))." Koethe, p. 95.

[11]Evidence for this line of argument can be found in passages such as: ". . . they [the holders of T_t] have a legitimate basis on which to reject the interpretaion of "true" as TRUE (SAT)--namely, that later speakers of the language (should there be any), or they themselves at some later time, probably could not accept it." Koethe, p. 95.

[12]See note 9.

[13]MMS, p. 22.

[14]Koethe, p. 95.

[15]Ibid., p. 95.

[16]Ibid., p. 95.

[17]See note 9.