On Putnam's Proof That We Are Not Brains-in-a-Vat
Author(s): Crispin Wright
Source: *Proceedings of the Aristotelian Society*, New Series, Vol. 92 (1992), pp. 67-94
Published by: Wiley on behalf of The Aristotelian Society
Stable URL: http://www.jstor.org/stable/4545146
Accessed: 22-11-2017 09:51 UTC

IV*—ON PUTNAM'S PROOF THAT WE ARE NOT BRAINS-IN-A-VAT[1]

by Crispin Wright

In the *Meditations* Descartes made sceptical play first with dreaming and then, more radically, with the idea that all conscious mental activity might be the creature of the notorious Demon. It is usually supposed that philosophy in this century has merely changed the example when its discussions proceed in terms of the nightmare of Brain-in-a-vathood. Here is how Hilary Putnam describes the nightmare in *Reason, Truth and History*:

> Imagine that [you] have been subjected to an operation by an evil scientist. Your brain has been removed from your body and placed in a vat of nutrients which keep [it] alive. The nerve-endings have been connected to a superscientific computer which causes [you] to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc.; but really all [you] are experiencing is the result of electronic impulses travelling from the computer to the nerve-endings. The computer is so clever that if [you] try to raise your hand, the feedback from the computer will cause [you] to 'see' and 'feel' the hand being raised. Moreover, by varying the programme, the evil scientist can cause [you] to 'experience' (or hallucinate) any situation or environment the evil scientist wishes. He can also obliterate the memory of the brain operation, so that [you] will seem to yourself to always have been in this environment. It can even seem to [you] that you are sitting and [hearing] these very

* Meeting of the Aristotelian Society, held in the Senior Common Room, Birkbeck College, London on Monday 18th November, 1991 at 8.15 p.m.

[1] This paper was originally presented at the Gifford Conference on the Philosophy of Hilary Putnam held at the University of St.Andrews in November 1990. A version of it will appear in the proceedings of that conference, to be published under the title *Reading Putnam* by Basil Blackwell and edited by Peter Clark and Bob Hale, to whom I am grateful for permission to read the present version to the Aristotelian Society.

> words about the amusing but quite absurd supposition that there is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients … .

However, Putnam goes on to modify and embellish the story in a couple of respects. The predicament is generalised, and the Evil Scientist drops out. He writes:

> Instead of having just one brain-in-a-vat, we can imagine that all human beings (perhaps all sentient beings) are brains-in-a-vat … . Of course, the evil scientist would have to be outside—or would he? Perhaps there is no evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems… . Let us suppose that the automatic machinery is programmed to give us all a *collective* hallucination, rather than a number of separate unrelated hallucinations. Thus, when I seem to myself to be talking to you, you seem to yourself to be hearing my words. Of course it is not the case that my words actually reach your ears—for you don't have (real) ears, nor do I have a real mouth and tongue. Rather, when I produce my words, what happens is that the efferent impulses travel from my brain to the computer, which both causes me to 'hear' my own voice uttering those words and 'feel' my tongue moving, etc., and causes you to 'hear' my words, 'see' me speaking, etc. In this case we are, in a sense, actually in communication. I am not mistaken about your real existence (only about the existence of your body and the 'external world', apart from brains). From a certain point of view, it doesn't even matter that the 'whole world' is collective hallucination; for you do, after all, really hear my words when I speak to you, even if the mechanism isn't what we suppose it to be.[2]

Why the changes? There is a superficial advantage. The standard version of the story which features the Evil Scientist and the brain operation, has the victim effectively incommunicado. In Putnam's version, by contrast, the existence of a *plurality* of

[2] *Reason, Truth and History* (Cambridge University Press, 1981) pp. 5–7.

communicative minds is not, at least *prima facie*, called into question. So we are free to discuss the issues arising without any existential absurdity, or so it seems. But of course the real significance of Putnam's modifications goes deeper—in fact, there are a number of ways in which they matter, which will emerge in due course.

The usual philosophical purpose of such a fantasy is to raise questions concerning the status of our beliefs about the material world. We strongly believe that we are in no such predicament as vat-hood. But how, unless we can back that belief up with hard evidence, is it anything more than a prejudice? And how, in that case, is any more flattering a verdict to be delivered on the ramified system of specific beliefs about the material world which we actually hold?

For much of what follows, I shall proceed as if the discussion in *Reason, Truth and History* is best received as addressed to the sceptical problem which such fantasies raise. However, I think that that is, ultimately, a mistaken interpretation. Putnam's real project is, as so often, to embarrass the metaphysical realist. But the achievement of his discussion in that direction will be easier to appraise after we have reviewed it in the context of epistemological scepticism.

II

How exactly do such fantasies wreak *general* sceptical damage? The answer might seem obvious. In intention at least, the fantasy will be so constructed as to be beyond defeat by any possible evidence. If it follows from its indefeasibility that I cannot reasonably be assured that the fantasy is false, then neither can I reasonably be assured that my relationship to the external world is not, in the way depicted, radically otherwise than I normally take it to be. And that, presumably, undercuts my right to assurance in all the great multiplicity of propositions which belong with the more congenial picture of the world, and of my interaction with it, which I actually hold.

But this rather loose train of thought misses an important distinction. The standard brain-in-a-vat fantasy is, whereas Putnam's is not, consistent with the *truth* of most of my beliefs about the material world. It may be that the purposes of the Evil Scientist do not require him to be a deceiver, and that most of the information I am fed is genuine. By contrast, Putnam's version of a world wholly constituted by a group of brains-in-a-vat and the attendant automatic machinery is already inconsistent with almost everything I actually believe about the physical universe. This discloses a more substantial consequence of Putnam's modifications to the standard story: an explicit sceptical argument built upon Putnam's version can be significantly simpler. Let 'R' express the epistemic notion—knowledge, warranted belief, or whatever—which the sceptical argument challenges. Then its first claim will be to the effect that

not R not [BIV],

i.e. that there is no knowing, or warrantedly believing, etc. that one is not a brain-in-a-vat. But reflect that if 'R' is transmissible across entailment, so will 'not R not' be. So given transmissibility, it follows that if the above claim is true, so is

not R not [P],

for any proposition, P, entailed by—so whose negation is inconsistent with—the original fantasy. If the fantasy is Putnam's, the negations of almost all my beliefs about the external world will thus be appropriate substituends for 'P', and it will follow directly that I cannot know, have warrant for, etc., any of those beliefs. If, on the other hand, the fantasy is the standard version, it is at least not possible so directly to conclude

not R not [P]

for most substitutions for 'P' of that kind. (Exceptions, of course, will be propositions concerning the physical condition and spatial deployment of my body.)

The difference, in other words, is this. A sceptical argument which works with Putnam's fantasy can directly transmit our (putative) lack of warranted assurance that the fantasy is false into a lack of warranted assurance that most of our ordinary beliefs about the material world are true. But no such direct transmission is possible if the argument works with the standard fantasy. Rather it will need to be argued in addition first that, even though their truth would be strictly consistent with the fantasy, my warrant for almost all my beliefs about the material world would still somehow be undermined by the truth of the fantasy; and, second, that, in order to be warranted in holding those beliefs, I require not merely that the fantasy be as a matter of fact false but some *assurance* that it is so. Argument for both points can be provided. But additional presuppositions have to be made. Putnam's version importantly simplifies the implicit presuppositions of the attendant sceptical paradox.[3]

As is familiar, Putnam argues that his version,

> ... although it violates no physical law, and is perfectly consistent with everything we have experienced, cannot possibly be true. *It cannot possibly be true*, because it is, in a certain way, self-defeating.[4]

The key to his argument is a simple comparison between the thought that I am a brain-in-a-vat and the thought that I do not exist: *viz*., that a necessary condition of my ability to think either thought is that it be false. At times Putnam writes as though the comparison is driven by a causal *theory* of reference—a theory according to which the reference of an expression would actually be determined by the obtaining of a favoured kind of causal relation between tokenings of it and the entity or kind of entities for which it stands. It is therefore important to realise, for anyone sceptical whether such a theory of reference could be correct, that none is strictly needed for Putnam's purpose. It will suffice if reference is merely, in appropriate cases, a *causally constrained* relation—that elephants, for instance, would not be the reference of 'elephant', as standardly used in English, if there were not appropriate causal relations between real elephants and tokenings of the term. Allowing that this much causality is necessary for reference in relevant cases is quite consistent with repudiating any suggestion that references generally can be identified by their satisfaction of certain causal conditions, (let alone that the concept of reference can be given a causal analysis).

[3] The indirect strategy of sceptical argument is the focus of attention in my 'Scepticism and Dreaming: Imploding the Demon,' *Mind* C (1991), pp. 87-116.

[4] *Reason, Truth and History* , p. 7.

The relatively conservative character of Putnam's appeal to causation is, of course, a strength of his discussion. The heart of his argument is summarised in the following passage:

> ... when the brain-in-a-vat (in the world where every sentient being is and always was a brain-in-a-vat) thinks 'there is a tree in front of me' his thought does not refer to actual trees. On some theories ... it might refer to trees in the image, or to the electronic impulses that cause tree experiences, or to the features of the program that are responsible for those electronic impulses. These theories are not ruled out ... [by the causal constraint] ... for there is a close causal connection between the use of the word 'tree' in vat English and the presence of trees in the image, the presence of electronic impulses of a certain kind, and the presence of certain features of the machine's program... .
>
> By the same argument, 'vat' refers to vats in the image in vat English or something related (electronic impulses or program features), but *certainly not to real vats* [my emphasis], since the use of 'vat' in vat English has no [relevant] causal connection to real vats... . Similarly, 'nutrient fluid' refers to a liquid in the image in vat English, or something related (electronic impulses or program features). It follows that if [the brains'] 'possible world' is really the actual one, and we are really the brains-in-a-vat, then what we now mean by 'we are brains-in-a-vat' is that *we are brains-in-a-vat in the image* or something of that kind (if we mean anything at all). But part of the hypothesis that we are brains-in-a-vat is that we aren't brains-in-a-vat in the image (i.e. what we are 'hallucinating' isn't that we are brains-in-a-vat). So, if we are brains-in-a-vat, then the sentence 'we are brains-in-a-vat' says something false (if it says anything). In short if we are brains-in-a-vat then 'we are brains-in-a-vat' is false. So it is (necessarily) false. [5]

The essential claim here is that tokenings of the expression 'brain-in-a-vat' in the thought of an envatted thinker, working in a

[5] *Reason, Truth and History* , pp. 14-15.

language—let's say: BIVese—lexically indistinguishable from English, would not sustain the right kind of relations to actual brains and vats to qualify as referring to brains-in-vats. Whatever, if anything, such tokenings did succeed in referring to, it would not be to brains-in-vats. But then, were we brains-in-vats, so Putnam's thought seems to be, the attempt to configure our true situation in thought would necessarily miss its mark—our thoughts would have as their truth-conditions, if they had any, some type of situation distinct from and incompatible with the real situation: our being brains-in-vats. So, Putnam concludes, the hypothesis is self-refuting, and necessarily false in consequence.

Putnam remarks that it took him a long time to convince himself that this exceedingly slippery train of thought is right. I have, myself, thought very different things about it at different times. It is only too easy, when one tries to make it fully explicit, to hit on formulations which invite good but inessential objections—objections to dispensable features imported by the particular formulation. But rather than spend time illustrating that, I shall immediately offer a formulation which I think is simplest and which best captures Putnam's intent. I shall claim that this argument is sound. But I do not think it fits all of Putnam's advertisement. In particular, it does not sustain the conclusion that, in the way we would like, the nightmare is refuted.

III

What are the suppositions which really drive Putnam's train of thought? There seem to be essentially two. One is the thought that *this* language—the one which we are using to reflect on these matters—permits correct, homophonic characterisation of the references, satisfaction-conditions and truth-conditions of meaningful expressions within it; in short, that this language *disquotes*. The second is that the language, or simulacrum of a language,[6] which would be operated by a brain-in-a-vat could not

[6] It is desirable, of course, to formulate the argument in a way which avoids the presupposition that brains-in-a-vat could *have* a language. The formulation which follows has also been guided by the desirability of avoiding subjunctive conditional claims of the form, 'If we were brains-in-a-vat, then [something would be so which isn't]', which inevitably seem question-begging in this context.

contain any expression referring to brains-in-a-vat. That suggests the following basic premisses:

(i) My language disquotes;

(ii) In BIVese, 'brain-in-a-vat' does not refer to brains-in-a-vat;

And now we may apparently proceed very directly. Thus

(iii) In my language, 'brain-in-a-vat' is a meaningful expression.

(iv) In my language, 'brain-in-a-vat' refers to brains-in-a-vat—from (i) and (iii).

Hence

(v) My language is not BIVese—from (ii) and (iv).

But

(vi) If I am a brain-in-a-vat, my language, if any, is BIVese—definition of BIVese.

So

(vii) I am not a brain-in-a-vat. QED.

This looks solid. But we need to negotiate two objections. The first concerns disquotation. Anthony Brueckner[7] canvasses a worry which, applied to the above formulation, would come to the thought that even the very modest degree of semantic externalism involved in supporting premiss (ii) will cancel the authority of the subject with respect to the content of expressions in her language. If I do not, at least before the argument is completed, know whether I am a brain-in-a-vat or not, then I don't know what if any kind of thing is so related to my tokenings of 'brain in a vat' to qualify as its reference. So I don't know what 'brain-in-a-vat' refers to, and am in no position, as at line (iv), to identify it by disquotation.

The proper reply to this, it seems to me, is to allow that there may well be a kind of identifying knowledge of content which, if semantic externalism is true, we do not possess. But if so, such identification is not a precondition of intelligent disquotation. Suppose the envatted brains do think determinate thoughts—that the components in their thought-symbolism are so related to real

[7] In his 'Brains in a Vat', *Journal of Philosophy*, LXXXIII, (1986), pp. 148–67.

items that they constitute BIVese as a genuine language, apt for the expression of truths and falsehoods. Then whatever that language is, it may presumably be correctly used in homophonic characterisations of reference and truth-conditions. Let it be that I do not know whether I am speaking English or BIVese; still I do know that, whatever 'snow' refers to in my language, I may identify its reference by *using* that very word. Isn't that enough to justify the disquotation?

I envisage the reply that, if that is as far as my knowledge goes, I am justified no further than in affirming merely that the instance of the disquotational scheme,

> 'Snow' refers to snow,

is a true *sentence*—that

> '"Snow" refers to snow' is true—

and that I am in no position to affirm *its* disquotation. But that is mere legislation. Consider a parallel with proper names. A subject's possession of identifying knowledge of the bearer of a name is no necessary condition for his legitimately using the name, or for others' ascribing to him thought-contents in whose specification they use the name. It is no solecism to say that a thinker knows that 'the Scarlet Pimpernel' denotes the Scarlet Pimpernel, but has no identifying knowledge of the Scarlet Pimpernel—does not know who the Scarlet Pimpernel is. It ought, in parallel, to be no solecism to allow that I know that 'Snow' refers to snow, even if, to whatever extent is enjoined by semantic externalism, I may not know *which* kind of stuff is snow or which thought the thought that 'Snow' refers to snow is. In order comprehendingly to disquote the sentence, '"Snow" refers to snow', I do not need, in any sense jeopardised by semantic externalism, to *identify* the thought that 'Snow' refers to snow. It is enough if, courtesy of the appropriate external circumstances, I *have* that thought, on the appropriate cue.

A related point may be made concerning subjects' authority for their own thought contents and the threat to it which semantic externalism has often, in recent literature, been thought to carry. Such authority does not depend on identification of the contents of one's attitudes, in whatever is the jeopardised sense. It is enough that my second-order beliefs—beliefs that I believe, or desire, or hope, etc., that P—be reliable: that it tends to be true that I believe, or desire, or hope, etc., that P just when I think that I do. And if that would be so but for semantic externalism, then it is quite unclear what difference semantic externalism could make. Unreliability would be a matter of *not* being disposed to believe that I believe, etc., that P just when in fact I do. But why should the consideration that the content of the belief that P is in part determined externally, by real-world relations which I may have no knowledge of, provide a reason for thinking that such a situation is more likely than it would be otherwise? Doesn't the opposing thought merely trade on overlooking that the content of my second-order beliefs will *likewise* be externally determined—that it will, as it were, co-vary with externally determined variation in the content of my first-order attitudes?

I suggest that there is no sound concern about the use made of disquotation in the argument, though there is much more to say about how externalism does impinge on self-knowledge. I'll come back to the matter.

The second objection worries about the role of the first-person in the argument. The thought seems compelling that, if the argument is valid, it ought to make no difference if its tokens of the first-personal pronoun and possessive are uniformly replaced by tokens of something which is for you, the assessor, a co-referring expression—say, 'CW'. But it does make a difference. Suppose you are given

(i) CW's language disquotes;

(ii) In BIVese, 'brain-in-a-vat' does not refer to brains-in-a-vat;

(iii) In CW's language, 'brain-in-a-vat' is a meaningful expression.

The argument then proceeds

(iv) In CW's language, 'brain-in-a-vat' refers to brains-in-a-vat—from (i) and (iii).

Hence

(v) CW's language is not BIVese—from (ii) and (iv).

But

(vi) If CW is a brain-in-a-vat, CW's language, if any, is BIVese.

So

(vii) CW is not a brain-in-a-vat.

The glaring problem is line (iv). Without supplementary information, you cannot validly infer anything from (i) and (iii) about how to *specify* what is the reference of 'brain-in-a-vat', as used in CW's language. All you can infer is that a specification in CW's language would be homophonic. That is the same thing as (iv) only if it is presupposed that your language—the language in which the argument is presented—is CW's.

Suppose your language isn't CW's—suppose in fact that CW is a brain-in-a-vat, the scenario differing from Putnam's only in that the vat, the contained brains and the attendant automatic machinery are situated in a vault in the British Museum. Then the falsity of the conclusion of the argument need pose no threat to the truth of its premisses. Let it be that CW's language—BIVese, on the present supposition—does indeed disquote and contain 'brain-in-a-vat' as a meaningful phrase. You may then infer that some disquotational sentence in BIVese which mentions and uses 'brain-in-a-vat' is true. But, without further information, that is all you may infer. There is nothing to contradict premiss (ii), that in BIVese, 'brain-in-a-vat' does not refer to brains-in-a-vat. So the argument, when formulated non-first-personally, is invalid.

But we should reply that if *all* that faults it is its suppression of the premiss that the language of its hero is the language in which it is formulated, then we ought to sustain the first-personal formulation. For isn't one effect of running the argument in the first person just to ensure that this premiss is true?

And that, it seems to me, is right. The first-personal formulation *is* valid. Moreover there seems no good objection to its premisses. Surely this language in which I am working does disquote; surely 'brain-in-a-vat' is one of its meaningful expressions; surely 'brains-in-a-vat' could not refer to brains-in-a-vat, whether BIVese is a language or not. So doesn't that do it ? What objection can there be to a valid argument whose premisses are recognised as true?

Consider the third-personal version which adds the following additional initial premiss

> (o) The following argument is formulated in CW's language, which coincides with that of you, the assessor, in the meaning assigned to all the expressions involved save possibly that assigned to 'brain-in-a-vat' and 'BIVese'.

Armed with this information you are now in position—assuming all expressions in the argument are used univocally—both to recognise that line (iv) is validly inferred from lines (i) and (iii), and that it is inconsistent with line (ii). Consequently you can recognise the argument as valid. What you are not able to do, if this is the limit of your information, is to ratify the argument as a valid derivation of the proposition that CW is not a brain-in-a-vat, since nothing in the stated premisses entitles you to a view about what the conclusion of the argument at line (vii) amounts to—it is, e.g., consistent with everything you know that 'brain-in-a-vat' in CW's language means: Frenchman, and 'BIVese': French.

What the first-personal version adds to the third-personal version augmented by (o) is, of course, exactly a putative closure of this gap: it is as if the words 'save possibly that assigned to "brain-in-a-vat" and "BIVese"' had been deleted from (o). But now, even this strengthened premise would not significantly change matters if it so happened that 'brain-in-a-vat' was an expression of my language which I did not understand, or only imperfectly understood, or if there was some indeterminacy about its content. And this is the crux. Earlier I argued that semantic externalism should not be viewed as generating any difficulty for disquotation or for knowledge of the contents of one's own attitudes. But I left it open that there might be a *kind* of identifying knowledge of one's own thought contents which, if semantic externalism is true, we do not after all possess. What, it seems to me, has driven those who have expended significant effort looking for invalid moves in the argument is a worry which is actually consistent with acknowledging both that the argument is valid and that its premisses are, plausibly, correct. The worry is whether fully exorcising the sceptical doubt which brain-in-a-vat examples and other similar stories raise would not require precisely the kind of identifying knowledge of content which semantic externalism itself proscribes. If that is right, then no argument which, like Putnam's, presupposes that content is a function, at least in part, of external factors—of the real relations holding between elements in our thought and the external world—can ever succeed in dispelling the basic intuitive concern.

We need, then, to address two questions. Is there a kind of identifying knowledge of content which we might have wanted to claim but which is called in question even by the very modest degree of semantic externalism involved in Putnam's argument ? And is it the case that finally putting the spectre of Brain-in-a-Vathood to rest would demand this kind of knowledge? I think the answer to the first question is 'yes'; the second matter is less straightforward, but eventually I shall suggest that, in a way, the spectre survives to haunt us still.

IV

It is part of the way we ordinarily think about self-consciousness that we regard the contents of a subject's contemporary propositional attitudes as something which, to use the standard term of art, they can *avow*—something about which their judgements are credited with a strong, though defeasible authority which does not rest on reasons or ‹vidence. No doubt it would be disturbing if plausible forms of semantic externalism called this conception into question. But, so far as I can see, there is, for the reason alluded to earlier, no threat. The perceived threat derives from the consideration that a subject has no such special, non-inferential authority in relation to external affairs; nor therefore, in particular, in relation to external affairs that enter into the determination of the content of her attitudes. The response was that since the contents of her second-order attitudes are determined *in tandem* with those of her first-order attitudes, externalism opens up no new way in which her actual psychology might diverge from her self-conception; whatever determines the content of the belief that P will *eo ipso* determine what belief it is that a subject who credits herself with the belief that P believes herself to possess.

Granting this is consistent, however, with recognising that a certain transparency of content *is* surrendered by externalism. And what this comes to is actually something perfectly definite. Externalism poses no threat to the *synchronic* identification of the contents of one's attitudes. But it does raise an issue about *diachronic* identification: specifically, about the identification of the contents of attitudes held in the past. For obviously, if what thought I express by tokening a sentence is in part determined by external factors of which I may be unaware, then I may also be unaware of sufficient of a change in those factors materially to affect what belief I express by a token of the same sentence on a subsequent occasion. So avowals of the form, 'I used to believe that P', or 'When I was a boy of six, I already believed that P', are opened by externalism to a new kind of defeasibility; sufficient external change of the right kind may defeat such a claim even when it is perfectly sincere and there is no question of a failing of memory.

Even so, it does not seem that much of a shake-up is engendered in our ordinary idea of the non-inferential authority of a subject's self-conception. The practice of granting authority to what a subject has to say about her former beliefs and desires can perfectly well rest on deep general contingencies—as so many of our practices do. Most people are good at recognising faces. And such recognition typically takes place spontaneously, without conscious inference. Clearly it is only against a background in which the world is not replete with look-alikes that we can possess such a skill. But our ordinary practice takes that fact for granted. So too, our ordinary practice may be viewed as taking it for granted that in the usual run of things semantically relevant external factors are not prone to change in reference-altering ways, still less to unmonitored such change. A picture associated with the practice of treating subjects as authoritative about former thoughts—the picture, roughly, of the contents of the mind as a kind of object, penetrated to the essence in ordinary self-knowledge—; this picture *is* a casualty of semantic externalism. But the reasonableness of the practice need not be impugned by the demise of the picture.

The answer to our first question, then, is that semantic externalism has no evident bearing on the authority with which we identify our thoughts; but that it has a conceivable bearing on the authority with which we *re*-identify them. The second question is accordingly:

would a finally satisfactory exorcism of the spectre of brain-in-a-vathood demand such re-identification of thoughts in a context in which its possibility is jeopardised?

It will help here to construct an example where a sceptical suggestion does resist exorcism by a Putnam-style argument for just this reason. Recall Twin-earth where all water-like stuff is XYZ but which otherwise resembles the Earth as an identical twin. But change the example to ensure that any instantiation of a natural kind found on Earth is matched by one of the same natural kind on Twin-earth, and conversely. What remains, then, is the duplication of objects and events. Suppose moreover that Twin-earth is not a counterfactual state of the Earth but an actual planet somewhere and that Earth and Twin-earth are quite unique in the similarity they bear to each other. And now let me be smitten with the paranoid fantasy that, at some point in my life, I will be—from my point of view—undetectably transferred there, never again to see family and friends, etc.—(it is a nice question why I should *mind*)—and never to learn what has happened.

Perhaps I can reassure myself by a judicious appeal to semantic externalism. The fantasy involves that once on Twin-earth, I will remain long enough—say 10 years—for the changed causation of my uses of the word 'Edinburgh', e.g., to effect a shift in its reference. If my fears materialise, then at some point by 'Edinburgh' I will no longer mean: Edinburgh. That suggests an argument of now familiar structure:

(i) My language disquotes;

(ii) In the language—Twenglish—of those who have been sufficiently long on Twin-earth, 'Edinburgh' does not refer to Edinburgh.

(iii) In my language, 'Edinburgh' is a meaningful expression.

(iv) In my language, 'Edinburgh' refers to Edinburgh—from (i) and (iii).

Hence

(v) My language is not Twenglish—from (ii) and (iv).

But

(vi) If I have spent the last 10 years on Twin-earth, my language is Twenglish.

So

(vii) I have not spent the last 10 years on Twin-earth. QED.

The reasoning is valid, and its premisses are true. And the putatively comforting reflection is not merely that I can run this argument now, to prove that so far all is well; but also that I will be able to dust the argument off and use it at any *future* time—so I know now that my fears will never be realised. My future on Earth is seemingly assured on purely semantic grounds.

Obviously, this is ineffectual. For an argument of just this form can be sound in the mouth of a Twenglish speaker. Reflect that, according to the modified Twin-earth story above, the semantics of English and Twenglish do not diverge except in the references assigned to proper names and other singular terms. So the only words in premiss (ii) whose meanings are liable to vary between English and Twenglish are 'Twenglish' itself, 'Twin-earth' and 'Edinburgh'. Now since the Twin-earth story does not just concern objects but it involves that all Earthly events, including socio-linguistic events, have Twin-earthly counterparts, it follows that whatever occurred on Earth to confer reference on Earthly uses of 'Twin-earth' will have been matched by counterpart occurrences on Twin-earth conferring reference on Twin-earthly uses of 'Twin-earth'. It isn't perhaps absolutely forced that the latter will consequently refer to Earth, but let us suppose that reference-fixing descriptions like 'the only planet in the universe which is a perfect twin of the Earth' have featured prominently in the sociology of the term on Earth. In that case, premiss (ii) will be true in the mouth of a Twenglish speaker, and uses of 'Twenglish', 'Twin-earth' and 'Edinburgh' in Twenglish will refer to English, Earth and Twin-Edinburgh respectively. Plainly, in this case, I *will* indeed be able to run the argument from true premisses at any future time; but this has no bearing whatever on the realisability of my fear.

We need not trouble over the exact formulation of the modest semantic externalism which drives the thought that the reference of my uses of 'Edinburgh' would change in due course, were a transfer to Twin-earth to take place. It is enough that replacing an object which plays a certain kind of role in the causation of one's use of a name by another which proceeds to play the same kind of role is taken to suffice, under the right circumstances and in due course, for the latter to supplant the former as the referent. Without this externalism, the argument could not get off the ground, for there would be no case for premiss (ii); it would be open to an objector to say, for instance, that, no matter how long I were to stay on Twin-earth after transfer, the reference of my uses of 'the Earth' would always remain to the Earth. But with the externalism in place, and under the circumstances envisaged, I will be in no position to detect the changes in reference which occur as a result of transfer; so in no position to know whether, for instance, the satisfaction-conditions of 'the only planet in the universe which is a perfect twin of the Earth' have changed in tandem with the reference of 'the Earth'. Using the argument against the paranoid fantasy requires that I be able to affirm that what I can now express by e.g. 'In 15 years time, I will not have spent the preceding 10 years on Twin-earth' will share its truth-conditions with what I can affirm in 15 years time by 'I have not spent the last 10 years on Twin-earth'. The argument entails that the latter will then be true. But, once I take the fear seriously, I must recognise that I will be in no position in 15 years to re-identify the thought then expressed as essentially that expressed now by 'In 15 years time, I will not have spent the preceding 10 years on Twin-earth'; and indeed, that if my fear has been realised, the truth of the future thought will be no fulfilment of the present thought. In short: the externalism which drives premiss (ii) also, in the circumstances of the example, hinders my identification now of the content of crucial claims made at future times; and the proof that those claims will be correct when made consequently gives me nothing I can use to allay my fears.

The question now is: does the proof that I am not a brain-in-a-vat lose its significance in the light of this reflection? Clearly it *would* do so if used to assail the standard version of the fantasy, featuring the Evil Scientist and envatting surgery carried out when I am already linguistically mature. For in that case there is a *type* of thought—that which I express by the words, 'I am not a brain-in-a-vat', at points in my life at which I have not yet been envatted—on which I can target my attention and desire afterwards, even though I cannot then express thoughts of that type by the same form of words. I can target it, for instance, as: the type of thought I would have expressed by the words, 'I am not a brain-in-a-vat', when I was first old enough to understand them. Well, *that* is the type of thought which I wish to be assured has no true instance; and it is, especially, the instance of that type apt to be rendered true or false by my present situation which I very much want to be true. But I cannot assure myself that it is so by running the proof that I am not a brain-in-a-vat, since the externalism needed to sustain an analogue of premiss (ii)—without which the argument cannot proceed—will also compromise my ability, as soon as I take the nightmare seriously, to identify the thought which I want to be true as the conclusion of the argument: the thought that I am not a brain-in-a-vat.

But of course it is different if we are concerned with Putnam's fantasy—another respect in which his changes matter. According to Putnam's version, we have *always* been envatted brains, tended by automatic machinery, in a universe where that is all there is. In order for the proof that I am not a brain-in-a-vat to miss its intended mark in the fashion indicated, we have to be able somehow to fix our thought on the intended mark—the proposition which we want the proof to disprove but which, in consequence of the operative externalism, we cannot certify whenever we want that it does disprove. But how is this to be done? There is no change in my status in Putnam's scenario, so no relevant semantic change: reference to the type of thought I would have expressed by the words, 'I am not a brain-in-a-vat', when I was first old enough to understand them will fix on nothing other than the type of proposition expressed by uses of 'I am not a brain-in-a-vat' at all times in my life.

The temptation may be to respond that there plainly are two possible types of thought to be considered: that which a normal human subject can establish by Putnam's proof and that which he could establish—prescinding from any worry about the possibility of envatted thought—were he a brain-in-a-vat. And what we need is an assurance that it is a thought of the former type which we establish when we run the proof. But the temptation, though understandable, is confused. In order for the proof really to concern a thought of the latter, unwanted type, I have to suppose that I am really a brain-in-a-vat as I run it. And that is a supposition which, according to the externalism governing the argument, I can make only if it is false. So there is all the assurance we could need!

'But surely a community of brains-in-a-vat could work through just these thoughts, and so convince themselves quite spuriously that they were not brains-in-a-vat?' No, they could not. They might work through these *words*, and soundly convince themselves of something. But only creatures which are not brains-in-a-vat can have these *thoughts*.

V

Putnam's argument, I suggest, is thus a transcendental argument which works. The question is, how much does it prove, and about what?

First, on scepticism. Say that a hypothesis H is *semantically auto-disruptive with respect to a language L* if and only if, were H true, some elements in the L-expression, S, of H would differ in meaning in such a way that S would no longer express H. And now define H as *absolutely* semantically auto-disruptive—*absaud*—if and only if for *any* expression, S, of H, in whatever language, if H were true some elements in S would so differ in meaning that S would no longer express H.

Granted that in order to entertain any particular hypothesis, some kind of recourse is necessary to a symbolic expression of it, it follows from the above characterisation that the thinkability of any absaud hypothesis requires its falsehood—for if the hypothesis were true, no symbolism would express it. There are two corollaries. First, that any reason to suppose that such a hypothesis is understood is reason to suppose it false. Second, in consequence, that no sceptical argument can reasonably be supposed cogent which satisfies both the following conditions:

(i) What is in fact an absaud hypothesis is claimed by one of its premisses to be beyond evidence for or against, and so beyond justified disbelief; but

(ii) Following the argument requires understanding that hypothesis.

For (ii), and the absaudity of the hypothesis, require that the argument can be followed only if the hypothesis is false; and reason to think the argument cogent obviously requires reason to think it can be followed. So reason to think the argument cogent requires reason to think the hypothesis false; that is, by (i), reason to think one of its premisses is untrue; that is, reason to think it is not cogent.

In these terms, Putnam's essential point is very simple: it is that the hypothesis that I'm a brain-in-a-vat, in the context of the relevant type of externalism and the further detail which he supplies, is an absaudity. For thinking it requires the use of symbols tokenings of which are causally linked to actual items in ways in which no tokenings by the brains-in-a-vat in his scenario can be. The thinkability of the hypothesis implies its falsehood, and it therefore poses no sceptical threat.

What apparently limits the epistemological significance of this result is that, as in effect already remarked, good sceptical arguments seem to have no need to make play with absaud hypotheses. Interestingly, the Cartesian Demon is presumably an absaudity. But the supposition that I am now dreaming, for instance, or the supposition that I was envatted yesterday, are not. The class of sceptical arguments in which we are interested all seek to exploit the putative first-person undetectability of some cognitively disabling state. What interest can it have if some examples of such states on which discussion has traditionally focused turn out to involve absaudity unless they *all* do? Otherwise, the conclusion should be merely that some traditional sceptical arguments employ inept examples. If our concern is with the general bearing of undetectable cognitive impairment on our scheme of putative knowledge and justified belief, then the failure of certain rather lurid (putative) illustrations of such impairment is of no consequence unless there is reason to think that such failure necessarily afflicts any illustration.

That is too fast. Earlier we remarked on a distinction between direct and indirect strategies of sceptical argument. The direct strategy, recall, trades on the evidence-transcendence of a hypothesis which, if true, would be immediately inconsistent with most of what we ordinarily believe about the relevant subject matter—in the present context, the material world. Its key thoughts are that if warranted belief is transmissible across entailment, then lack of warrant to disbelieve will be likewise; that evidence-transcendence enjoins lack of warrant to disbelieve; and, hence, that we lack warrant to disbelieve the negations of, i.e. to hold, most of our beliefs about the material world. The indirect strategy, by contrast, works with a hypothesis which is again presumed evidence-transcendent, at least for its subject, but whose truth is not actually inconsistent with the class of beliefs that the sceptical argument is concerned with. The strategy is to argue, rather, that those beliefs, even if their truth does not require the falsity of the hypothesis, are not appropriately *warranted* unless the hypothesis is false—that their being warranted requires that they have a pedigree which is inconsistent with the truth of the hypothesis; that warrants have to be recognisable; and hence that warrant for such beliefs requires justification—allegedly impossible—for our taking the hypothesis to be false. This is the way it goes, for instance, with the dreaming hypothesis. The sceptic charges that no subject at any time has warrant for the belief that he is not then dreaming. To be sure, one can dream what is actually taking place. But beliefs formed on the basis of the apparent perceptual experience that features within a dream do not have the pedigree of actual perceptual warrant. Justification of those of my beliefs which I take to be perceptually warranted by my current experience would thus depend on justification for the belief that I am not now dreaming which, according to the sceptical contention, is impossible.

There is no space now to review the detail of the indirect strategy in any depth. Suffice it to say that it has, as it were, significantly many more moving parts than the direct strategy and is to that extent more likely to break down.[8] The striking point is therefore that, at least as far as the hypotheses so far mentioned are concerned, the division between the absaud and the non-absaud corresponds exactly to that between those of them which are apt to

[8] And does indeed do so—see my *op.cit.* note 3.

subserve the direct strategy of sceptical argument and those which are available only for the purposes of the indirect strategy. If I have always been a brain-in-a-vat, with the background detail as in Putnam's scenario, it follows directly that almost all my beliefs about the material world are false. If Descartes' Demon has practised massive deception on me as far as the material world is concerned, then again, it follows directly that almost all my beliefs about the material world are false. But if the thought is merely that, for all I can show to the contrary, I might now be dreaming, then it is at least consistent to suppose that what I am now dreaming is actually true, even if, *qua* dreamer, I am thereby deprived of any good grounds for believing it. Likewise, if the story is that I was envatted last night in such a way as to give rise to no gross discontinuity in my experience, then it is again quite consistent to suppose that beliefs which I have formed since envatment should be true.

The following thought accordingly occurs: has Putnam perhaps at least shown that the *direct* strategy of sceptical argument necessarily aborts? That is, is it the case that any hypothesis at the service of the direct strategy—a hypothesis of sufficient logical strength to be directly inconsistent with most of what we believe about e.g. the material world—has to be an absaudity? It would certainly constitute significant epistemological progress if this were so.

Obviously there are, from the sceptical point of view, risks in making play with hypotheses of such strength. If, for instance, the hypothesis works like Putnam's version of the brain-in-a-vat hypothesis, by cancelling the existence of most of the material world we believe in, then it will be committed to some deviant story about the aetiology of the uses of material world vocabulary that feature in our thought, and may very well clumsily impinge, in consequence, on the semantics of symbols that feature in its own expression. But while—at least in the presence of the kind of semantic externalism which has informed our whole discussion—it is difficult to see how such a hypothesis could fail to be semantically disruptive, it's hard to be clear *a priori* that it would have to be semantically *auto-disruptive*.

On reflection, though, mere semantic disruption is enough. Avoiding absaudity is in any case no guarantee, if you want to be a sceptic, of avoiding the trouble given by Putnam's argument. For the argument can easily be adapted to the refutation of *any* semantically disruptive hypotheses, whether or not the disruption extends to its own expression. Consider a sceptical hypothesis, H, whose semantic disruption is confined just to a single predicate. Then

(i) My language disquotes;

(ii) In a language, L, differing from mine, if at all, only in respects dictated by the truth of H, 'F' does not refer to F's.

(iii) In my language, 'F' is a meaningful expression.

(iv) In my language, 'F' refers to F's—from (i) and (iii).

Hence

(v) My language is not L—from (ii) and (iv).

But

(vi) If H is true, my language is L—from (ii).

So

(vii) H is not true. QED.

So now it would seem that the direct strategy of sceptical argument against our material world beliefs has a very tall order to fill: somehow or other, it must devise a hypothesis which is both directly inconsistent with most of our material world beliefs and *wholly semantically conservative*—wholly consistent with the satisfaction of whatever causal conditions are involved in determining that all our material world expressions have the references which they actually have. One might well scent the possibility of an *a priori* demonstration that those are incompatible requirements.

But the Sceptic is nothing if not resourceful. Reflect that one sure-fire way to avoid semantical disruption is to leave the past alone, so that whatever causal constraints have been observed in the socio-linguistic history which has determined the actual references of our expressions continue to be observed in the story. Reflect too that one way of ensuring that most of our beliefs about the material world are false is to have most of what we take to be the material universe not exist. Then the obvious way for the direct-strategy material world sceptic to have at least some of his cake while eating the rest is, accordingly, to propose a hypothesis whereby the material world was, until recently, much as we had always taken it to be, but has recently largely ceased to be. Such a story might be, for instance, one in which the evil scientist envats a large number of us, sets up the attendant automatic machinery, and then, by a fitting nemesis, accidentally destroys himself along with the rest of the material universe. Or let the Evil Demon have lain doggo up until yesterday, when he destroyed the material world in which I believe and first assumed his classic Cartesian role. Such a hypothesis will subserve the direct strategy when targeted against our beliefs concerning the present and future states of the material world, but—provided the cataclysm is supposed to have been of some recency—should not be semantically disruptive at all. It remains to be settled whether Putnam's argument can indeed subvert all direct-strategy forms of material world scepticism which aspire to a certain generality; but it is clear that the generality involved must at least embrace our beliefs about the past.[9]

VI

I conclude that the epistemological significance of Putnam's proof is limited at best. But earlier I suggested that, by its author's intention at least, the real bearing of the proof might be less

[9] Does this point to a loophole for 'the Sceptic' to exploit? Do these Limited Direct Strategy (LDS) arguments slip between both Putnam's argument and the 'Implosive' counter-argument developed against the Indirect Strategy in my *op.cit.* note 3? Well, reflect that there is nothing to stop our exporting into the Indirect setting the sort of sceptical hypotheses which LDS arguments will feature; hence, if the counterargument of 'Imploding the Demon' is sound, an issue will be raised in any case about the inference from the (putative) impossibility of accumulating evidence against such hypotheses to the claim that there is no warrant for their denial. But once that inference is placed *sub judice*, then LDS Scepticism is going to have trouble making a case for its premisses unless it can independently find fault either with elements in the apparatus deployed in the Indirect strategy of which the Direct strategy has no need but which the Implosion exploits, or with the Implosion itself. So at worst there is no *extant* LDS sceptical paradox—the sceptical work, if it can indeed be done, is still to do. I hope to discuss the matter more fully on another occasion.

epistemological than metaphysical. I shall finish by pursuing this a little.

The view of the world which Putnam calls metaphysical realism is certainly nothing very precise. It involves thinking of the world as set over against thought in such a way that it is only by courtesy of a deeply contingent harmony, or felicity, that we succeed, if we do, in forming an overall picture of the world which, at least in its basics, is correct. This is what commits the metaphysical realist to the possibility that even an ideal theory might be false or seriously incomplete. And the same kind of thinking surfaces in the idea that the world comes pre-jointed, as it were, into real kinds, quite independently of any classificatory activity of ours. Once one thinks of the world in that way, one is presumably committed to the bare possibility of conceptual creatures naturally so constituted as *not* to be prone to form concepts which reflect the real kinds that there are. The real character of the world and its constituents would thus elude both the cognition and the comprehension of such creatures.

Putnam's brains-in-a-vat are exactly such creatures: minds doomed by the character of their interaction with the world they inhabit, and by the nature of that world, not to have the concepts they need in order to be able to capture in thought that world's most fundamental features and the nature of their relationship with it. The modifications which Putnam effected in the standard version of the brain-in-a-vat story thus enable it to serve as a kind of parable for the kind of cognitive predicament which metaphysical realism is committed to regarding as a real possibility. And the following argumentative strategy is thereby opened up. Show that the brain-in-a-vat fantasy is false in a way which depends only on features essential to its being the right kind of parable and you have provided an argument which will show that *any* such parable is false. But possibilities are things that might be realised, and when they are, there has to be some *specific* way in which they are realised—some specific description of the state of affairs which realises them. Metaphysical realism is committed to the possibility of a certain kind of dislocation, or uncrossable divide between reality and our cognitive activity. If that possibility were realised, there would, accordingly, have to be some correct specific account of the way in which it was realised. And that is just to say that something like the brain-in-a-vat story would have to be true. The details might naturally be very different; but the essential overall structure would be the same. It would be an account whereby, despite the apparent cognitive richness of our lives, we were somehow so situated as not to be enabled to arrive at the concepts which fundamentally depicted the character of the real world and the nature of our interaction with it.

Possession of a general method for refuting any such account would thus apparently put us in position to convict metaphysical realism of something akin to Ω-inconsistency. An Ω-inconsistent system of arithmetic, you recall, is one which, for some arithmetic predicate, F, both contains a proof that there is an x such that not Fx and proofs of each statement of the form, Fn, 'n' a numeral. Simple inconsistency is avoided only because the recognition that each Fn is provable is not accomplished *via* means formalised within the system. The metaphysical realist would be committed to claiming, correspondingly, the irrefutability of the hypothesis that we are in a cognitive predicament of a certain very general sort, even though for each *specific* description of such a situation, the claim that we are in *that* situation would be open to definite refutation.

The unclarities about what is essential to metaphysical realism may seem to make it hard, at first blush, to assess the dialectical promise of this line of criticism. But in fact I think it's quite clear that it will fail. The difficulty is that Putnam's proof does not represent a general method for disproving *any* specific version of the relevant kind of possibility; at best, it represents a general method for disproving any specific version *which we can understand*. Any impression to the contrary involves temporarily forgetting that the falsity of the brain-in-a-vat hypothesis is established conditionally on its thinkability. But the sort of dislocation whose possibility is arguably implicit in metaphysical realism does not involve that its victims can conceptualise their predicament; quite to the contrary—their predicament consists in part precisely in the fact that they are debarred from arriving at the concepts necessary to capture the most fundamental features of their world and their place in it. There is nothing akin to Ω-inconsistency in the thought that a certain type of situation is possible—that we ourselves may be in such a situation—although any instance of the type whose specific characteristics we can conceptualise can be demonstrated not to obtain. The only conclusion licensed is merely that if the type is realised, we will not be able to understand the specific form in which it is so.

And this I think is the real basis of the dissatisfaction that so many have felt with Putnam's proof. The trouble is that we easily slip into inept formulations of it. The dissatisfaction is ineptly formulated, for instance, when presented as the idea that there are two relevant types of thought which the sentence, 'We are not brains-in-a-vat', might express, depending whether we are brains-in-a-vat or not, and that we need an assurance, which Putnam's proof cannot give us, that the one we succeed in refuting is the right one. When the doubt is expressed like that, we have all the assurance we need just in the proof that we are not brains-in-vats. But the real spectre to be exorcised concerns the idea of a thought *standing behind* our thought that we are not brains-in-a-vat, in just the way that our thought that they *are* mere brains-in-a-vat would stand behind the thought—could they indeed think anything—of actual brains-in-a-vat that 'We are not brains-in-a-vat'. The spectre is that of a thought whose truth would make a mockery of mankind and its place in nature, just as our true thought that they are merely brains-in-a-vat makes a mockery of the 'cognitive' activity of the envatted brains. What we should really like would be an assurance that there is no such true thought: an assurance not just that most of what we think is actually true—for semantic externalism might well deliver that result for the brains-in-a-vat—but that we are on to the right categories in terms of which to depict the most general features of the world and our place in it; and can be reasonably assured that we are thinking about such matters in the right general kind of way.

But of course, if there were such a true thought, standing behind us as it were, it would no more be available to us than the thought that they are merely brains-in-a-vat would be available to the envatted brains. It is, indeed, unclear whether reflective philosophy can possibly deliver the assurance we would like to have. But if it can, it will not be by argumentative manipulation of intelligible, specific descriptions of the kind of predicament, epitomised by the brains-in-a-vat, to whose possibility metaphysical realism is committed. By the same token, a proper disenchantment with the metaphysical realist outlook must be motivated in a different way.[10]

Department of Logic and Metaphysics
The University
St. Andrews
Fife
Scotland KY16 9AL

[10] Though no-one who has heard or read this paper has yet suggested so, it is quite possible that others have published similar thoughts on the subject, and I owe an apology to anyone who has. There is now, of course, a very large body of work on Putnam's proof and the limited time I had in which to compose a paper for the Gifford Conference meant I had no opportunity for a properly scholarly survey and indeed had little option but to ignore the secondary literature altogether. An exception was Anthony Brueckner's paper referred to above. The immediate stimulus to think again about the topic was provided by another, unpublished paper of Brueckner's, 'Semantic Answers to Skepticism', read at the University of Michigan in Fall 1989; and by an intervention of Allan Gibbard in that discussion. I have been much helped by discussions with Allan Gibbard, Bob Hale, Stephen Yablo and Hilary Putnam, and by colloquia at Edinburgh, the Oxford Ockham Society, and King's College, London.