

Brains in a Vat, Subjectivity, and the Causal Theory of Reference

Kirk Ludwig
Department of Philosophy
University of Florida
Gainesville, FL 32611-8545

1. Introduction

In the first chapter of *Reason, Truth and History*,¹ Putnam argued that it is not *epistemically* possible that we are brains in a vat (of a certain sort). If his argument is correct, and can be extended in certain ways, then it seems that we can lay to rest the traditional skeptical worry that most or all of our beliefs about the external world are false. Putnam's argument has two parts. The first is an argument for a theory of reference² according to which we cannot refer to an object or a type of object unless we have had a certain sort of causal interaction with it. The second part argues from this theory to the conclusion that we can know that we are not brains in a vat.

In this paper I will argue that Putnam's argument to show that we cannot be brains in a vat is unsuccessful. However, the flaw is not in the argument from the theory of reference to the conclusion

¹ Hilary Putnam, *Reason, Truth and History*, (Cambridge: Harvard University Press, 1979). Henceforth 'RTH'. The position that Putnam advances in this first chapter is one that in later chapters of RTH he abandons in favor of the position that he calls 'internal realism'. He represents the arguments he gives in chapter 1 as a problem posed for the 'external realist', who assumes the possibility of a God's eye point of view. I will not be concerned with the later development of Putnam's view.

² I mean 'reference', as I think Putnam does, to be taken broadly here, to mean roughly what sort of thing or things a term purports to be about. I talk for the most part in terms of reference rather than representation in this paper to follow more closely Putnam's terminology. See the end of section 8 for a discussion of Putnam's use.

that we are not brains in a vat, as has often been alleged. Most of Putnam's critics have charged that *even if* the causal theory of reference were correct, this would not secure for us knowledge that we were not brains in a vat, because an essential assumption of the argument is that we can know what language we speak or what we mean by our words, and this assumption is undermined or question begging once we accept the causal theory of reference.³ I will argue that Putnam's critics have not substantiated this charge. The causal theory of reference does not *by itself* entail that we cannot know what our words mean. The mistake is in the argument *for* the theory of reference. I will argue that it rests ultimately on a false picture of the mind, one which leads us to a false picture of the relation between the contents of our thoughts their subjectivity, their character from the point of view of introspection. Moreover, in virtue of this, if the picture of the mind implicit in Putnam's *argument* were correct, then the conclusion that Putnam's critics reach would be warranted. Thus, the argument is faced with a dilemma. If Putnam's argument for this theory of reference were correct, then we could not know the meaning of our words or, more generally, the contents of our thoughts, for the picture of the mind implicit in the argument divorces the only possible basis of knowledge of meaning and content from the logical determinants of meanings and content. If we reject this picture of the mind, while this does not show that the causal theory of reference is false, Putnam's argument for it collapses. In either case, we fail to show that we can know that we are not brains in a vat of the sort Putnam describes.

In section 2, I present Putnam's argument for his theory of reference. It is important to do this carefully because my criticism will hinge upon the way Putnam frames his thought experiments. In section 3, I present the argument from that theory of reference to the conclusion that we cannot be brains in a vat. Sections 4 and 5 explain and rebut the charge that Putnam's argument fails because if

³ See note 13 for references. For an exception, see Thomas Tymoczko, "In Defense of Putnam's Brains," *Philosophical Studies* 57 (November 1989): 281-298. See also Tyler Burge's more general defense of relational theories of thought content against this charge in "Individualism and Self-Knowledge," *Journal of Philosophy* 85 (November 1988): 649-644. Although I can agree with part of Burge's conclusion, in my opinion his defense fails because it relies on an extremely impoverished view of the nature and scope of self-knowledge. See Paul Boghossian's "Content and Self-Knowledge" *Philosophical Topics*, 17 (Spring 1989): 5-26, for a criticism of Burge's conception of self-knowledge.

his *theory of reference* were correct, we couldn't know the meanings of our words. Section 6 examines and sets straight a confusion that creeps into Putnam's exposition of his argument. Section 7 shows that Putnam is after all open to the charge that his account, in *conjunction* with certain assumptions he makes in his argument, which do not enter into the *statement* of the theory, shows that we cannot know what we mean or think, and that what must be done to repair this undermines his argument for his theory of reference. Section 8 examines the sources of the mistake in the argument. Section 9 is a brief conclusion.

2. The argument for the causal theory of reference

Consider an ant crawling on a beach, which traces a line in the sand that crosses and recrosses itself so as to produce what could be taken to be a drawing of Winston Churchill. Despite the ant's tracing out a line on the beach that resembles a drawing or picture of Churchill, it does not represent or refer to Churchill. We can see it *as* a picture of Churchill, but it is not. Even if the line the ant has traced is physically type-identical to a picture of Churchill that someone has traced in the sand on that beach, it is still not a picture of Churchill. This illustration shows that pictures do not represent in virtue of any special sort of similarity or resemblance to the objects that they represent. If this is true of pictures, then *a fortiori* it is true of words considered as inscriptions or sounds (or physical signs of any sort). For these need not bear even a resemblance to what they refer to or represent. Clearly, we could run through the considerations above with a tracing in the sand of 'Churchill'. Physical objects, then, do not represent or refer to things intrinsically.

Putnam argues that what is true of pictures and words considered as physical signs is *also* true of mental images and words, and in fact mental entities of all sorts. "What is important to realize is that what goes for physical pictures also goes for mental images, and for mental representations in general" (RTH, p. 3). It is natural to think that what makes the difference between a physical sign or picture that represents or refers and one that doesn't is that when a sign refers to something it is because of some special sort of relation it bears to someone's thoughts. If someone draws a picture of Churchill in the

sand on the beach, intending it to be a picture of Churchill, then it is a picture of Churchill, even if it is a very poor one.⁴ And it is natural to think that whereas physical signs only derivatively refer to or represent things, a person's thoughts or images refer to things intrinsically. Putnam objects to this on two grounds, one theoretical, and one based on a series of thought experiments in which we are to judge that when someone has a mental image, e.g., or 'hears' some words in his head, perhaps accompanied by some feeling of understanding, it does not follow that he is referring to or representing anything at all.

The theoretical ground is that to assert that mental states refer intrinsically is to advocate what should be called a magical theory of reference. ("Mental representations no more have a necessary connection with what they represent than physical representations do. The contrary supposition is a survival of magical thinking" (RTH, p. 3).) It is a magical theory of reference because it offers no way to understand how reference is possible, or in virtue of what reference succeeds. It is the antithesis of a scientific account of reference or representation.⁵ In effect, a magical theory of reference is no theory of reference at all, but rather a refusal to give a theory. So if we are to give a theory at all, we cannot stop at saying mental entities just do refer to things intrinsically. We must say (informatively) in virtue of what they refer. If we undertake the project of giving a theory of reference, we must deny that mental entities refer intrinsically.

Putnam's first thought experiment asks us to consider a faraway planet on which a race of humanoids live who have never seen or heard of trees because none exist on their world. One day one of these humanoids comes across a piece of paper on which some paint has been spilled and has formed an image visually exactly like a picture of a tree. "Suppose," Putnam says, "one of them has a mental image exactly like one of my mental images of a tree as a result of having seen the picture. His

⁴ Though no doubt if it is intended specifically as a *picture* of Churchill, some resemblance is a condition for success; a straight line or a stick figure wouldn't do.

⁵ In later work Putnam apparently gives up this constraint, for he says, "I mean to deny that there is some scientifically describable 'nature' that all cases of 'reference' in general, or of 'meaning' in general, or of 'intentionality' in general possess." See *Representation and Reality*, (Cambridge: MIT Press, 1988), 2.

mental image is not a *representation of a tree*. It is only a representation of the strange object (whatever it is) that the mysterious picture represents” (RTH, p. 4). The same thing, Putnam says, is true of mental ‘words’. If, for example, I memorize a speech in Japanese that I hear on the radio, and repeat it parrot-like to myself, *sotto voce*, or even just in my head, the words spoken or going through my mind don’t represent or refer to anything at all, even if all of this (perhaps under hypnosis) is accompanied by a feeling of understanding. In fact, if this is right, it is clear that someone could both think words that were, in some language, about trees and have appropriate mental images without referring to or representing anything at all about trees. To put this another way, it is possible that everything going through someone’s conscious mind could be just the same as what is going through mine when I am thinking about and imagining trees without his knowing or thinking anything at all about trees. Putnam supposes it would be very unlikely that this would ever happen, but says that “if it did happen, it would be a striking demonstration of an important conceptual truth; that even a large and complex system of representations,⁶ both verbal and visual, still does not have an *intrinsic* built-in, magical connection with what it represents--a connection independent of how it was caused and what the dispositions of the speaker or thinker are” (RTH, p. 5). As Putnam sums it up: “Thought words and mental pictures do not *intrinsically* represent what they are about” (RTH, p. 5).

Putnam supplements these thought experiments with one intended to show that not only are complex systems of words and images not sufficient to determine what they are about, but that even if we add to this “rules⁷ deciding what words may appropriately be produced in certain contexts--even if we consider, in computer jargon, *programs for using words*--unless those programs themselves refer

⁶ We should perhaps not call this large system of mental states a ‘system of representations’, since they do not represent anything when picked out the way Putnam wants to pick them out. They are rather what’s left over after we subtract representational content, however we are to imagine doing this.

⁷ These can’t be rules which themselves make reference to things in the environment, e.g., a rule such as ‘Say “There are cows” if and only if there are cows’ where this is definitely in, say, English. For then the rule would be sufficient to determine what someone was referring to. The rule must give conditions for uttering words syntactically described which specify conditions that do not presuppose any reference to external objects.

to something extra-linguistic there is still no determinate reference that those words possess” (RTH, p. 10). To show this Putnam asks us to consider a Turing Test for Reference. This is a variation of the test Alan Turing suggested to determine whether a computer is conscious.⁸ Turing suggested that to give the question whether a computer is conscious any sense one would have to devise some test success at which would count as showing that something was or was not conscious, and he argued that the best test for whether a computer was conscious would be to see whether it could successfully play the Imitation Game. In Turing’s version of the Imitation Game a human being tries to tell of two interlocutors, one of whom is another human being, and one a computer, which is which. In order not to allow extraneous matters to intrude, the conversations are conducted using typewriters, so that, e.g., what an interlocutor looks like (all wires and metal surfaces) will not influence the decision. If the computer cannot be reliably discriminated from the human interlocutor, then, according to the Turing Test, it is conscious.

A question that can be raised about the Turing Test is whether it is a test of what it claims to be, that is, whether we are justified in saying that a computer that passes the Turing Test *is* conscious (or that we have as much warrant for saying it is conscious as for saying the human interlocutor is conscious). Putnam’s aim in imagining a Turing Test for Reference is to raise a parallel question. Suppose someone suggested that if an interlocutor, say a computer, successfully played the Imitation Game, that would show that it was referring with its words. Would he be right? More specifically, imagine that we are playing the Imitation Game with a machine which can pass the test but “has no sense organs ... and no motor organs.” Moreover, “not only does the machine lack electronic eyes and ears, etc., but ... there are no provisions in the machine’s program, the program for playing the Imitation Game, for incorporating inputs from such sense organs, or for controlling a body” (RTH, p. 10). If this machine successfully played the Imitation Game, could we say that its words referred to anything? Putnam says the answer is No.

⁸ Alan Turing, “Computing Machinery and Intelligence,” *Mind* 59 (October 1950): 433-460.

[I]t seems evident that we cannot and should not attribute reference to such a device. It is true that the machine can discourse beautifully about, say, the scenery in New England. But it could not recognize an apple tree or an apple, a mountain or a cow, a field or a steeple, if it were in front of one. (RTH, p. 10)

We should not attribute reference to the device because none of its sentences “is at all connected to the real world” (RTH, p. 10). In the absence of any connection, we naturally judge that it can’t be referring to the real world. In contrast,

Our talk of apples and fields is intimately connected with our *non-verbal* transactions with apples and fields. There are ‘language entry rules’ which take us from experiences of apples to such utterances as ‘I see an apple’, and ‘language exit rules’ which take us from decisions expressed in linguistic form (‘I am going to buy some apples’) to actions other than speaking. Lacking either language entry rules or language exit rules, there is no reason to regard the conversation of the machine ... as more than syntactic play. (RTH, p. 11)

The machine is utterly insensitive to the existence of apples, or anything other than itself. “That is why the machine cannot be regarded as referring at all” (RTH, p. 12).

What’s missing, Putnam says, is appropriate causal contact with the objects the machine is putatively referring to. The point carries over to the referring abilities of human beings. To avoid embracing a magical theory of reference we must hold that reference is possible only if there is some sort of contact, a “non-verbal transaction,” between a person and what he refers to. Putnam suggests at one point we might rely on “noetic rays,” but dismisses this as crazy (RTH, p. 51). About the only alternative, the only naturalistic or non-magical relation that could do the trick, seems to be some sort of causal contact between us and the objects we refer to. A crucial component in the description of our thought experiments was that the words or thoughts under examination had no causal connection, or only a very weak one, with what they were putatively about or represented or referred to. A natural

explanation for our judgments in these cases, then, is that there was no causal contact with the objects the subjects might have been thought to refer to. This conclusion is perfectly general: it rules out that any sort of mental entity, not just mental images or some other introspectible quality, can intrinsically refer.

The position so far reached, then, is that a necessary condition for reference to, or representation of, a certain thing or kind of thing is causal contact of a certain sort with that thing or things of that kind. As we have seen, Putnam supports this conclusion with two considerations. One is the theoretical consideration that the position that attributes intrinsic powers of reference to mental states is unscientific (i.e., magical). The other is a series of thought experiments designed to elicit the conclusion that in the absence of causal contact with putative objects of reference or representation neither resemblance between a mental image and the object, nor phenomenological identity with the state of one who would be referring to the object, nor a complete set of rules for using words is sufficient for reference to the object or objects.

3. The argument to the conclusion that we are not brains in a vat

This conclusion is the basis of Putnam's argument that we can know that we are not brains in a vat of a certain sort. Specifically, the possibility that Putnam asks us to imagine is that "all human beings (perhaps all sentient beings) are brains in a vat" (RTH, p. 6). In the version of the brain in the vat hypothesis Putnam asks us to consider, there is no evil scientist outside of the vat, "the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems" (RTH, p. 6). The machinery operates so as to produce a collective hallucination, so that the many brains in the vat have experiences just like the experiences we have. To ensure complete causal isolation from ordinary objects, we are to imagine that this has always and will always be so.⁹

⁹ For convenience, I will use the phrase 'brains in a vat' to abbreviate this fuller description of the possibility Putnam imagines, and likewise "brains in a vat" to abbreviate the quotation name of the sentence describing this possibility.

Putnam asks us to “Suppose this whole story were actually true,” and then asks “Could we, if we were brains in a vat in this way, *say* or *think* that we were?” His answer is No, because, although “the supposition that we are actually brains in a vat ... violates no physical law, and is perfectly consistent with everything we have experienced, [it] cannot possibly be true. *It cannot possibly be true*, because it is, in a certain way, self-refuting” (RTH, p. 7).

It is self-refuting, according to Putnam, in the same sense in which the sentence “I do not exist” is self-refuting whenever anyone thinks or says it. That is to say, it is self-refuting because “the supposition that the thesis is entertained or enunciated ... implies its falsity” (RTH, pp. 7-8). Putnam says that the supposition that we are brains in a vat (of the sort described above) has this property. “If we can consider whether it is true or false, then it is not true ... Hence it is not true” (RTH, p. 8).

This is so despite the fact that

The humans in that possible world [the one in which they are brains in a vat of the sort described above] have exactly the same experiences that *we* do. They think the same thoughts we do (at least, the same words, images, thought-forms, etc., go through their minds). (RTH, p. 8)

But when put this way, the claim seems very puzzling. How can it possibly be true? The key is supposed to be that, although these brains in a vat can ‘say’ or ‘think’ anything we can, as far as their subjective experiences go, they cannot refer to the things that we can, because in their situation they lack the appropriate sort of causal contact with the things that we refer to. In particular, their words ‘brains’ and ‘vat’ lack the right kind of causal contact with brains and vats to be referring to brains and vats.

Let’s suppose that this is true, and that if they cannot refer to what we do by their words ‘brains’ and ‘vat’, they do not (could not) mean (in whatever the ordinary sense of that word is) what we do by ‘brains’ and ‘vat’. In this case their sentence, ‘We are brains in a vat’, does not express what our sentence, ‘We are brains in a vat’, expresses. Let us assume further that when a

brain-in-a-vat-worlder says 'we' it does manage to pick out the members of its linguistic community, or at least itself. (Although Putnam does not make these assumptions explicit, I think we are justified in attributing them to him on the grounds that these make best sense of his argument.) Then, in saying 'We are brains in a vat', the brain-in-a-vat-worlder is asserting (if anything) something of itself and other members of its linguistic community (if any) other than that it is (or they are) a brain (or brains) in a vat. Suppose further than none of their words is appropriately causally connected with brains or vats so that they can refer to (or think about) them.¹⁰ Then they simply cannot express the thought that we express when we think or say 'We are brains in a vat'. Putnam puts this in the following way:

Once we see that the *qualitative similarity* (amounting, if you like, to qualitative identity) between the thoughts of the brains in a vat and the thoughts of someone in the actual world by no means implies sameness of reference, it is not hard to see that there is no basis at all for regarding the brain in a vat as referring to external things. (RTH, pp. 13-14)

We assume, of course, that the brains in the vat in this world are intelligent and conscious; "it would seem absurd" to deny this. Nonetheless they can't be referring to or representing or thinking about the same things we do with their words, for they are not in appropriate causal contact with them. If this is right, then Putnam has proved that these brains in a vat cannot think or say that they are brains in a vat. They cannot think or say it for the same reason that someone whose language does not contain words expressing what we mean by 'neutrino' and 'rest mass' cannot think or say that neutrinos have no rest mass.

¹⁰ The assumption is not entirely trivial. Consider their words, 'We are people in the world'. The word 'people', if they causally interact, is connected with brains, and their word 'world' is causally connected with the vat they are in. An extensional translation of their sentence, 'We are people in the world', then, might well be 'We are brains in the vat'. If this is an incorrect translation it will be because the role of 'people' and 'world' in their language is significantly different from the role of 'brain' and 'vat' in ours.

If the brains in the vat in this world are conscious and intelligent, then it seems reasonable to think of them as capable of referring to things, even if they are not the same things we refer to. It is, I believe, because Putnam assumes that the brains in the vat would be intelligent and conscious, and that reference is necessary for thought, that he assumes that the brains in the vat would be referring to something despite being in the vat.¹¹ “On some theories that we shall discuss,” he says, the statement “There is a tree in front of me” as uttered by a brain in the vat “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” (RTH, p. 14).

This gives us a stronger conclusion than merely that brains in a vat of a certain sort can’t think that they are brains in a vat. We are assuming that they are referring to things, and that it is a condition on their referring to things that they be in causal contact with them. But of course not just any sort of causal contact will do. Our theory must be non-magical, so it must make sense of why one word rather than another refers to a particular sort of thing. We won’t be able to say what special connections hold between words and objects for a brain in a vat by looking at its causal relations with its environment at an instant. What determines the content of some particular thought of a brain in a vat about its present environment must be what *regularly* causes it or would cause it to have that sort of thought.¹² So the kinds of causal relations to its environment that determine the content of a brain in a vat’s thoughts are the lawlike relations between its thoughts and things in its environment. (Putnam does not explicitly draw this conclusion, but I think it is implicit in the reason he gives for saying the brains in a vat would not be referring to the vat by their words because there is no “special connection between the use of the *particular* word ‘vat’ and vats” (RTH p. 15).)¹³

¹¹ A consequence of this assumption is that the failure of a Turing test for reference entails that it fails as a test for consciousness as well.

¹² We must think here of thoughts as something like word-forms that go through our heads, which are specifiable independently of content.

¹³ This theme is explicit in the work of causal theorists of content such as Fred Dretske (*Knowledge and the Flow of Information*, (Cambridge: MIT Press, 1981), *Explaining Behavior* (Cambridge: MIT Press, 1988)), Jerry Fodor (*Psychosemantics* (Cambridge: MIT Press, 1987)), and

A consequence of this is that if a brain in the vat's causal environment is relatively stable, most of the time when it thinks, e.g., 'There's a tree in front of me', what it thinks is true. For what fixes the references of its terms is what usually causes it (or would cause it) to utter or think them when it has an indexical thought about its immediate environment. That means that when it usually thinks or utters 'There's a tree in front of me' it is referring to the thing that its word 'tree' refers to when whatever *that* is stands in the relation to *it* that is expressed by its words 'in front of'. This goes for all of its sentences.

Putnam does not draw this conclusion from his reflections, but it seems to me to be a natural extension of this line of thought. It's important also if the argument is to show not merely that we know that we are not brains in a vat, but that we know in addition that most of our beliefs are true.¹⁴ We need this if we want a general defense against skepticism about the external world. For it is not enough to show that we cannot be brains in a vat if what we want to know is that we don't go massively wrong about the world around us. Being brains in a vat is only one way to go wrong.

In arguing to the conclusion that we are not brains in a vat, Putnam focuses on the suggestion that a brain in the vat would mean by 'we are brains in a vat' that "*we are brains in a vat in the image or something of that kind*" (RTH, p. 15). He argues that if this theory is true, then we cannot be brains in a vat. How do we get from the claim that a theory of this sort is true, to the claim that we are not brains in such a vat?

Donald Davidson ("A Coherence Theory of Truth and Knowledge," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore (Basil Blackwell: New York, 1986)).

¹⁴ Cf. Donald Davidson, "A Coherence Theory of Truth and Knowledge," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore (New York: Basil Blackwell, 1986) pp. 307-319.

In explaining this final step, Putnam gives the following sketch of the argument.¹⁵ If the theory is true, then, since the brains in the vat are by hypothesis brains in a vat, and these truth conditions are incompatible with their being brains in a vat in the image, the brains in a vat think or say something false when they think or say ‘we are brains in a vat’. “It follows,” Putnam says, “that if their ‘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” (RTH, p. 15). But if we are brains in a vat, we aren’t brains in a vat in the image. Therefore:

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. (RTH, p. 15)

It is not, on this view, physically impossible that we or beings like us be brains in a vat (of the sort described here). For it is not physically impossible for there to be a world where there are beings relevantly similar to us who are such brains in a vat. But it is logically impossible that in the actual world we are what we mean by ‘brains in a vat’. It must always be a counterfactual possibility, because of the way in which reference and representation is fixed. Possible languages and possible worlds get matched so that there is no possible world in which ‘is a brain in a vat’, where this expression has an analogous role to its role in our language, comes out true of any being in that world in its language. Yet, beings in one possible world w (or the actual world) might have been, in w^* , what they in w call ‘brains in a vat’, though in that case they would speak a different language than they do in w . So, when Putnam claims that he shows that we cannot possibly be brains in a vat, he is not claiming that it is not possible that we might have been such brains in a vat, but only that it is not possible that we actually are.¹⁶

¹⁵ The argument that follows differs subtly but importantly from the one outlined above, according to which “If we can consider whether it [the statement that we are brains in a vat] is true or false, then it is false.” We will come back to this below.

¹⁶ It is important to remember the qualifications on the situation ‘brains in a vat’ picks out. Putnam’s argument is directed against the possibility that we have always been and will always be

This concludes my presentation and extension of Putnam's argument that we can know we are not brains in a vat. In the next two sections, I take up the charge that Putnam's argument fails because if his theory were correct, we could not know what we mean by our words. In the section following these I return to a difficulty in the canonical formulation of Putnam's argument we have just examined.

4. Brueckner's objection

Anthony Brueckner¹⁷ has argued that although there is a valid argument from Putnam's theory of reference to the conclusion that our sentence 'We are brains in a vat' is false, that does not show that we know we are not brains in a vat. It fails to show that we are not brains in a vat because the theory of reference that is its basis gives rise to another sort of skepticism, which is, if anything, worse than the sort we set out to cure. If the argument is good, then the application to the brain in a vat possibility shows that we don't know the meanings of our own words, because we do not know whether we speak English or vat-English (or any of the indefinite number of varieties generated by other possibilities). And unless we know whether we speak English or vat-English, we cannot conclude from the falsity of 'We are brains in a vat' that we are not brains in a vat.

brains in a vat, not against the possibility, for example, that we have all been brains in a vat for the past ten years, prior to which we were embodied. The force of Putnam's anti-skeptical argument has sometimes been criticized on these grounds. Yet if he establishes that we can know we are not brains in a vat of the sort he describes, he shows that radical skepticism about the external world, the view that we can know nothing beyond the contents of our own minds and necessary truths, is false. That is quite a large enough accomplishment.

Still, there is a puzzle here. For if skepticism about the external world radically misrepresents our epistemic position with respect to the external world, if philosophical skepticism is an illusion, then a successful refutation of skepticism should restore to us our full epistemic innocence. To the extent to which Putnam's conclusion in this argument fails to do this, we may doubt that it has touched upon what is central to skepticism, why we feel its pull, why we should resist it. If that is right, then a solution to the skeptical problem may bypass altogether considerations of how the meanings of our words get fixed. Indeed, Putnam would probably agree with this: his distinction between the internal and external realist is an attempt to isolate a deeper problem in the skeptical position.

¹⁷ "Brains in a Vat," *The Journal of Philosophy*, (March 1986): 148-167. Henceforth 'BIV'.

This objection has been made repeatedly to Putnam's argument.¹⁸ I will discuss Brueckner's version of it for its clarity and thoroughness. But I will argue that the objection fails. Putnam's theory about how the meanings of our words are determined does not by itself entail that we fail to know them. I will also argue, though, that Putnam's theory in conjunction with other assumptions he makes does have the consequence that we cannot know either what we mean by our words or what we think, and that in giving up those assumptions, he must give up his argument for his theory of reference.

Brueckner sees Putnam's argument as attacking a particular sort of argument against skepticism that appeals to our inability to *rule out* the logical possibility that we are brains in a vat. Since Putnam does not argue, however, that it is logically impossible that we be brains in a vat, the argument cannot proceed simply by denying that it is logically possible that we are brains in a vat. Instead, Brueckner sees Putnam as arguing that whether or not we are brains in a vat, we can legitimately conclude we are not. And it is in this that Brueckner sees the difficulty for Putnam looming.

Brueckner offers the following interpretation of Putnam's argument (BIV, p. 154). (Brueckner uses the theory Putnam mentions according to which the brains in the vat refer to vats-in-the-image with 'vat', and so on, and takes 'vats-in-the-image' to pick out sense impressions of vats. 'BIV' is Brueckner's abbreviation of Putnam's description of the possible world he argues cannot be actual.)

Argument E

¹⁸ See, for example, Jane MacIntyre's "Putnam's Brains," *Analysis* 44.2 (March 1984): 58-61; Susan Feldman's "Refutation of Dogmatism: Putnam's Brains in Vats," *Southern Journal of Philosophy* 22 (Fall 1984): 323-330; Peter Smith's "Could We be Brains in a Vat?," *Canadian Journal of Philosophy* XIV (March 1984): 115-123; James Stephens and Lilly-Marlene Russow's "Brains in Vats and the Internalist Perspective," *Australasian Journal of Philosophy* 63.2 (June 1985): 205-212; Paul Coppock's "Putnam's Transcendental Argument," *Pacific Philosophical Quarterly* 68.1 (March 1987): 14-29; Gary Iseminger, *Analysis* 48.4 (October 1988): 190-195.

- (1) Either I am a BIV (speaking vat-English) or I am a non-BIV (speaking English).¹⁹
- (2) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.
- (3) If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV.
- (4) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are false. [(2),(3)]
- (5) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.
- (6) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are false. [(5)]
- (7) My utterances of 'I am a BIV' are false. [(1),(4),(6)]

Let's consider the first problem Brueckner raises. It seems reasonable to assume that

(T) My utterances of 'I am a BIV' are true iff I am a BIV

is true for both an English and vat-English speaker, since it is guaranteed to be true if the object-language and meta-language are the same. Now consider

(5) If I am a non-BIV (speaking English), then my utterances of 'I am a BIV' are true iff I am a BIV.

Given (T) we should also have

(8) If I am a BIV (speaking vat-English), then my utterances of 'I am a BIV' are true iff I am a BIV.

But when we combine this with

¹⁹ It is worth noting that although being a BIV or a non-BIV are exhaustive alternatives, speaking English or vat-English are not. This interpretation of Putnam's argument relies on taking speaking English and vat-English to be exhaustive alternatives.

(3) If I am a BIV (speaking vat-English), then I do not have sense impressions as of being a BIV

we find that (8) and (2) “give incompatible truth conditions for my utterances of ‘I am a BIV’ on condition that I am a BIV” (BIV, p. 157). To see this, suppose that I am a BIV. By (8) ‘I am a BIV’ is true. By (2), then, I have sense impressions as of being a BIV. But by (3) I do not have sense impressions as of being a BIV. Contradiction.

This argument, however, isn’t valid. It relies on an equivocation. (8) gives the truth conditions on the right hand side of the biconditional in vat-English, while (2) and (3) are in English. Brueckner brings up this objection not because he thinks it is a valid argument, but to raise the issue of what language argument E is being given in. To the equivocation charge, Brueckner protests that

If I am allowed to assume that I am speaking English rather than vat-English, then I am allowed to assume that I am not a BIV. In that case, the argument E is of no interest. If I do not assume that the argument is being given in English, though, the problem of evaluating the argument becomes quite bizarre. (BIV, p. 158)

Let’s assume for the moment that the first conditional in this passage is true, and that if I assume that I am speaking English, then argument E is of no interest. Then if the argument is going to be of any interest it must be one that we can be sure is valid independently of our knowing whether we are speaking English or vat-English. Brueckner assumes that it is valid in English. Is it valid if it is expressed in vat-English?

Brueckner suggests the following as the interpretation of the argument in vat-English, taking ‘sense impression’ in vat-English to refer to sense impressions.

(1’) Either I have sense impressions as of being a BIV or I do not have sense impressions as of being a BIV.

(2') If I have sense impressions as of being a BIV, then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.

(3') If I have sense impressions as of being a BIV, then I do not have sense impressions as of being a BIV.

(4') If I have sense impressions as of being a BIV, then my utterances of 'I am a BIV' are false [(2'),(3')].

(5') If I do not have sense impressions as of being a BIV, then my utterances of 'I am a BIV' are true iff I have sense impressions as of being a BIV.

(6') If I do not have sense impressions as of being a BIV, then my utterances of 'I am a BIV' are false. [(5')]

(7') My utterances of 'I am a BIV' are false. [(1'),(4'),(6')]²⁰

Only (2'), (3'), and (5') present a possible problem. Brueckner says that (2') and (3') are vacuously true because they have a "common necessarily false antecedent" (BIV, p. 161). If we assume that I am speaking vat-English, then it seems nothing would count as my having sense impressions as of being a BIV, for we have described the case so that the brains in the vat have the same experiences we do, and we never do, or could have experiences as of being a brain in a vat. It's hard to see, in fact, what would count. (5'), however, is false if the conditional is read as being stronger than a material conditional, because if, e.g., I am not a BIV, and do not have sense impressions as of being a BIV, then the truth conditions for 'I am a BIV' are that I am a BIV, and not that I have sense impressions as of being a BIV. (We are reading the biconditional as stronger than truth functional as well, if we are interested in truth conditions.) That is to say, the conditional is not true in all possible

²⁰ There are difficulties here that we will skip over. Are we to think of the words here as being the English interpretation of the vat-English words? But then it would be in English after all, and not in vat-English, and then we would be assuming that we are speaking English. Presumably also this is not how the BIV thinks of itself. Presumably there is no translation of 'BIV' into vat-English at all. According to Putnam the BIV can't even think it. So how could he say it? And what of the 'BIV' in the specification of the kind of sense impression? Is that vat-English or English?

worlds. But this doesn't matter for the present purpose. We are assuming that I speak vat-English, so the question is, with that assumption, is the premise true? Since if the antecedent is true, on the assumption that I am speaking vat-English, the consequent gives the correct truth conditions for 'I am a BIV', it is true in vat-English. So the conclusion follows. We can recast the argument in this form.

- (i) If I am speaking English, then my sentence 'I am a BIV' is false.
- (ii) If I am speaking vat-English (or some such variant), then my sentence 'I am a BIV' is false.
- (iii) Therefore, my sentence 'I am a BIV' is false.²¹

We conclude from (T) that I am not a BIV.

However, according to Brueckner, although it looks as if we've got a good argument, the victory over the skeptic is only apparent. "There is," he says, "a severe limit to the anti-skeptical force of our argument."

If I do not know whether I am speaking English or vat-English, then I cannot apply (T) to my own utterances of 'I am a BIV' as a step toward the conclusion that I know that I am not a BIV and hence am speaking English. ... since I do not know whether I am speaking English or vat-English, I do not know whether the truth conditions of my utterances of 'I am a BIV' are the strange ones specified in premise (2) or rather the disquotational ones specified in premise (5). (BIV, pp. 164-5)

Brueckner's objection seems to be that although I can conclude that 'I am a brain in a vat' is false, and so by (T) that I am not a brain in a vat (understanding this to be expressed in whatever language I speak), this does me no good if what I want to know is that I am not what is expressed in English by the sentence 'Ludwig is a brain in a vat', for I could know this only if I knew what language I was

²¹ For the conclusion to follow, we have to assume that speaking English or vat-English are the only alternatives.

speaking, and I could know this only if I already knew I was not a brain in a vat. Brueckner compares our position to that of someone who knows that a certain sentence, 'Omega is not a regular cardinal', is true, but does not understand the sentence. He can conclude from this and the corresponding T-sentence that Omega is not a regular cardinal. But he still does not know what proposition that expresses. Though he can say that Omega is not a regular cardinal, using the sentence 'Omega is not a regular cardinal' to express a proposition, in a certain sense he still does not know which proposition he is expressing. So, in fact, he does not know that Omega is not a regular cardinal. The same goes, on this view, for my knowledge that I am not a brain in a vat. Brueckner puts it this way: "The current problem facing our anti-skeptical argument ... is that it at best affords knowledge that a certain sentence expresses a false proposition, whereas the intended sort of refutation of skepticism depends upon the availability of knowledge that a certain proposition is false--the proposition that I am a BIV" (BIV, p. 165). "The anti-skeptical strategy reconstructed herein fails in the end because it engenders a sort of skepticism about meaning or propositional content" (BIV, p. 166). The sentence 'I am a BIV' expresses different propositions in English and vat-English, but unless I can know whether I am speaking English or vat-English, I cannot know what proposition it expresses. "All I can claim is the metalinguistic knowledge that a certain sentence expresses a false proposition, rather than the object-language knowledge that I am not a brain in a vat" (BIV, p. 167).

5. Can we know what we mean?

Brueckner's claim that Putnam's theory does not have the anti-skeptical force Putnam intended it to have rests on the assumption that if Putnam's theory of reference were correct, we would not know whether we are speaking English or vat-English (or any variant that arises from being embedded in the world in different ways). I am not sure why Brueckner thinks this is true. At one point he says that if I know that I am speaking English, then I already know that I am not a brain in the vat. And at another point he says that one is entitled to the assumption that one means by 'I am a brain in the vat' what ordinary human beings mean by that "only if I am entitled to assume that I am a normal human

being speaking English rather than a BIV speaking vat-English” (BIV, p. 160). And he objects that, “This must be *shown* by an anti-skeptical argument, not assumed in advance” (BIV, p. 160). Of course, it *is* supposed to be a result of the argument that I am not a brain in the vat. And if being a normal human being is not being a brain in a vat, then that must be shown by the argument. But it is not clear why knowing that one speaks English should be thought in the same way to be something that must be shown by the argument, rather than an assumption of it.

One reason Brueckner might assume that this must be shown by the skeptical argument is that he thinks Putnam did not intend to assume that we were speaking English in giving his argument. This is suggested by his remark that if we could assume that we were speaking English, we would know we were not brains in a vat, and the argument he had reconstructed would be “of no interest.” But if the reconstructed argument would in that case have been of no interest, we would still have had a valid argument against the possibility that we were brains in the vat, namely:

(P1) If we speak English, then we are not brains in the vat.

(P2) We speak English.

(C) Therefore, we are not brains in the vat.

Brueckner accepts (P1). If he accepted (P2), he would have to acknowledge that Putnam’s theory of reference yields a valid argument that shows we are not brains in the vat, which does not require that we consider whether it is valid in English and vat-English. Then going on to consider whether it is valid in vat-English would be a pointless exercise. So I think Brueckner’s thinking that Putnam intended that we not assume we are speaking English is not sufficient to explain why he goes on to construct an argument for the BIV, and that we must assume that Brueckner thinks that we do not know—at least if Putnam’s theory is correct—that we speak English unless first and independently we know that (C) is true. This is indicated by the way Brueckner states the first premise of the argument in English.

Why would Brueckner think we can’t know we are speaking English unless we know we are not brains in a vat? One line of thought that might lead one to the rejection of the claim that we know

(P2) is this: By Putnam's argument we know that if we are brains in a vat, then we are not speaking English, but instead vat-English. Conversely, if we are not brains in a vat (or otherwise relevantly differently embedded in the world), we are speaking English. Suppose it is in doubt whether we are speaking English or vat-English. Clearly, with what we have before us here, we could establish one or the other only if we already knew that we were or were not brains in a vat, as that is understood in English. But, then, how do we know we *are* speaking English?

This question apparently has, however, a simple answer. We know we are speaking English because 'English' is our word for *the language we speak*. We pick out the language we speak indexically, so there is no trouble about how we know what language we speak, and there is, really, at bottom, no other way to pick out the referent of 'English'. That we speak English is never in question, and is not called into question simply by establishing that if we were brains in a vat, we would not be speaking English. If we had all been born Frenchmen, we would not be speaking English. But no one would suggest we face a problem in knowing whether we speak English unless we have already established that we were not all born Frenchmen. To put this more generally, if we spoke some language other than English, then we would not be speaking English. But this is no obstacle to knowing that we speak English. This is true even if some other language has in it the word 'English' that is used by speakers of that language to refer to their language. We could dispense with the word 'English', in fact, in favor of the phrase 'our language'.

But the worry Brueckner has about Putnam's argument is not best put in terms of our knowing whether or not we speak English or vat-English. It is a worry about our knowing the meanings of our words. Of course we know that we speak our language, and that language is of course English (not vat-English, which we know is something else, since it is introduced by us to refer to a language which we can infer from Putnam's argument we do not speak). But do we know what the words in our language mean? If we did not, we could not know, from knowing that our sentence 'We are BIVs' is false, that we are not BIVs. But the claim that we could not know what we mean given Putnam's theory requires an argument. We have been given no more reason to believe that we could not know

the meanings of our words if Putnam's theory were correct than that we could not know that we were speaking English.

It will help to see why Brueckner's objection fails to consider more generally the nature of Putnam's response to skepticism and Brueckner's counter argument. We can best understand Putnam's argument against the epistemic possibility that we have always been brains in a vat as proceeding in two stages. First, we argue that if our words 'We are brains in a vat' mean that we are brains in a vat, we are not brains in a vat. Second, we conclude from this, on the basis of our *prior* knowledge that our words 'We are brains in a vat' mean that we are brains in a vat, that we are not brains in a vat. More abstractly, the strategy is to show that there is a logical connection between certain *external* facts, facts our knowledge of which was originally in question, and certain *internal* facts, facts our knowledge of which is not in question or under suspicion. The external fact is that we are not brains in a vat, and the internal fact is that our words 'We are brains in a vat' mean we are brains in a vat.²²

The form of Brueckner's response to this is that if Putnam's theory of what determines meaning is correct, then the meanings of our words become external facts. If the meanings of our words were external facts, then we could not assume we knew the meanings of our words without begging the question against the skeptic, who argues we know no external facts at all. Putnam's argument would be powerless against skepticism. At best we would know that a certain sentence, 'We are brains in a vat', is false. But we would not know what was expressed by that sentence, and so we would fail to have knowledge of the world around us.

I have defended Putnam against Brueckner's charge by arguing that Brueckner has simply assumed without argument that knowing the meanings of our words, if Putnam's theory were correct, would beg the question against the skeptic, would be tantamount to already knowing something about

²² Putnam does not put his argument this way in the apparently canonical form summarized at the end of section 3. As articulated there, the argument is flawed, as I explain in the next section. However, I think the way I have put the argument in the preceding paragraph best expresses the spirit of Putnam's argument, especially in his initial characterization of it. It should be clear that I do not think Brueckner's interpretation is the best way to understand it.

the world around us. On the face of it, knowledge of the meanings of our words is a paradigm of knowledge of internal facts. Skeptics about the external world typically do not question whether we know what we mean by our words. Indeed, it seems to be a presupposition of the skeptic's investigation, of any investigation, that we do know the meanings of our words. If we don't know the meanings of our words, we don't know what thoughts we express with them. For if we did know what thoughts we express with our words, we could know what we mean by them. If we can't know what our thoughts are, it's hard to see how we could even begin to investigate something. The suggestion that we don't know what we mean or think is mind boggling. That we do is as secure as anything is for us.²³

The charge that Putnam's theory deprives us of knowledge of the meanings of our words is a powerful one, because it amounts to the charge that Putnam's account is not only false but incoherent. How are we supposed to imagine the dilemma that Brueckner sees facing us when we conclude we can't know what a sentence of ours expresses? Is it that we wonder whether it expresses *this* rather than *that*? That can't be right. For that would imply that we could think both the thoughts in question. But in either case we could think only one. And if we are seriously imagining that we don't know what proposition our sentence expresses, if we can in some sense still think either thought in question, we cannot know that we do. If this is true of this sentence, then it is true of every sentence that depends on the nature of our environment for its meaning. The argument itself then becomes opaque to us, employing as it does a great many words that refer to external objects.

What our words mean can't be external facts. So any theory that implies that they are is false. But to substantiate the charge that Putnam's theory makes facts about meanings external facts, one must show that Putnam's account divorces the meanings of our words from the only thing our knowledge of them could rest on, or rests our knowledge of their meanings on something we could claim to know only if we had already shown skepticism to be false.

²³ This is not to say we can always give an adequate account in other words of what the meanings of our words are, or that this would be what such knowledge consisted in. This is one of the lessons, if we needed to be taught it, of Wittgenstein's remarks on rule following.

Why should one think this? Let me suggest two arguments that one might give to show that Putnam's theory makes facts about meanings external facts.

(i) The fact that I am not a brain in a vat, if it is a fact, is clearly an external fact. But on Putnam's view, that my words 'I am a brain in a vat' mean that I am a brain in a vat is true only if I am not a brain in a vat. And in fact the statement 'My words "I am a brain in a vat" mean that I am a brain in a vat' is logically equivalent to some, probably disjunctive statement about things in space. Since statements about things in space express external facts, so, on Putnam's account, do statements about the meanings of my words.

I don't think this argument is good. It is no response to an argument against skepticism to say that it must beg the question because if true it would show that we can know what the skeptic denies we know. But that is the effect of this argument. It does not undermine just Putnam's argument, but any argument that attempts to show that something we know entails some fact that the skeptic denies we know. Thus this objection has nothing to do with Putnam's argument in particular. It is not the causal theory of reference that is the culprit. It is that the argument purports to show that skepticism is false. The objection relies on these two assumptions. (a) If Putnam's theory is correct, then the fact that I am a brain in a vat will still be seen to be an external fact. (b) If some proposition is logically entailed by another, then if it expresses an external fact, so does the entailing proposition. The proper response is to observe that, on the one hand, if we accept (b), since we do know the meanings of our words, and could not if they were external facts, we should deny (a), while on the other hand, if we accept (a), then since we do know the meanings of our words, and could not if they were external facts, we should deny (b).

(ii) One version of the type of account Putnam accepts holds that the meanings of our words and contents of our thoughts are logically determined by the past pattern of our causal interaction with our environment. As I argued above, this seems a natural extension of Putnam's arguments. What our words mean is determined by what, in basic cases,²⁴ has usually caused us to apply them to things

²⁴ Of course, we learn many referring words with distant or even no causal contact with their objects, but such learning, on this view, is itself dependent on more direct conditioning of other words

around us in the past. My word 'desk', e.g., means desk, rather than diode, because I have in the past been, for the most part, in causal contact with desks when thinking indexical 'desk'-thoughts about my immediate environment, and not with diodes.

A key point is that the determinants of our thoughts are external things and states of affairs. To know what our words mean, we have to know what the determinants of their meanings are. That is to say, on the present view, to know what our words mean, we have to know what the past history of our causal interaction with the world has been, what sorts of things have caused us to utter 'desk'-sentences and have 'desk'-thoughts. But that's already to know something about the world around us. And whether we could know something of that sort was what was in question. So on this kind of account, we can know what our words mean only if we already know things about the world around us. If there was an initial problem about how this is possible, Putnam's conclusions can't solve it.

There are at least two ways to read the crucial "to know what our words mean, we have to know what the past history of our causal interaction with the world has been." On the first reading, to know what our words mean, we have to go out and investigate the world, find out what sorts of things have caused us to utter sentences of certain sorts and have thoughts of certain sorts, non-Intentionally described. We'd have to figure out what our own words meant in the same way someone else would. If this were what knowing the meanings of our words took, then it seems clear that, first, in assuming we can do this we would be begging the question against the skeptic, and that, second, we really couldn't get started, since we wouldn't know what we were thinking about the world until we'd finished the project, and we can't undertake the project unless we can know what our thoughts about the world are.

On the second reading, it is not that we have to discover what the causes of our utterances and thoughts are to know what their meanings and contents are, but that *in* knowing their meanings and contents, we know facts about our past causal interaction with the world. The second of these readings does not justify the conclusion that the question has been begged against the skeptic. Since we do

by what they are about. For convenience I will omit the qualification below.

know the meanings of our words, if we show that to know the meanings of our words is to already know something about the external world, then we will have shown that we know something about the external world. It is an adequate response to skepticism about some domain of facts to show that we know some facts in that domain.

The first reading is important because it shows what Putnam must hold to use this account of how meaning gets fixed to argue that we do or can have knowledge of things around us. The first reading claims that, on the causal account of the determination of meaning, an individual would have to figure out what his own words meant in the same way someone else would figure out what his words meant. It is clear that we don't and couldn't figure out what our own words meant in this way. The causal theorist must allow that the individual knows his own meanings in a way no one else does, that he has a perspective on what his own words mean (and the contents of his thoughts) different from that of someone observing his interaction with his environment, and so does not have to use the same procedure to determine what he thinks. If he did have to use that procedure, then he couldn't know the meanings of his words or the contents of his own thoughts.

For all that has been said so far, there is nothing to stop Putnam from claiming that one knows the meanings of one's own words and the contents one's own thoughts in a way different from the way someone observing one's interaction with the world would. If he can legitimately make that claim, then his theory of how meanings and thought contents are determined does not undermine our knowledge of the meanings of our words and contents of our thoughts.

The claim that there are two perspectives on anyone's thoughts and meanings also helps us to see that Putnam need not accept the second of the readings above. What Putnam's argument shows, if it is good, is that there is a logical connection between the meanings of one's words and what one's environment is or has been. For this to provide any sort of ground for knowledge of the world, he needs to assume that we know the meanings of our words. But we can know the meanings of our words independently of, in a way different from, how we know facts about the external world. Knowing facts about the external world then may involve an inference *from* what our meanings are. So it need not be that *in* knowing the meanings of our words we know things about the external world,

even if given that we know the meanings of our words, we know or can know things about the external world.

6. A problem in Putnam's explanation of his argument

The objection we've just examined fails because it fails to substantiate the charge that we don't know whether or not we are speaking English unless we know whether or not we are brains in a vat (as expressed in English). But I think it does point to a difficulty in the way Putnam has put his argument. Putnam sums up his argument by saying "If we are brains in a vat, then 'we are brains in a vat' is false." This looks as if it is a sentence of the form

If p, then 'p' is false.

where 'p' is replaced by the same sentence in the same language in each of its appearances. We can conclude from the fact that 'p' is false that not-p. So it looks as if Putnam is giving us a conclusion we can state in this form

If p, then not-p.

From which it follows that (necessarily) not-p. But this can't be the form of the argument, because in the conditional that Putnam affirms the truth conditions of the antecedent are the ones it has when interpreted in English, while the reason the sentence quoted in the consequent is false is that its truth conditions are given relative to vat-English. So the sentence quoted in the consequent is not interpreted in the same way as the one used as the antecedent. From the falsity of the sentence quoted in the consequent, interpreted relative to vat-English, we can conclude nothing directly about the truth or falsity of the antecedent, interpreted relative to English. But what we want to know is whether the sentence interpreted relative to English is true or false.

Let's see whether we can restate the argument along the lines Putnam does taking into account that the conditional is expressed in English while the sentence declared to be false is in vat-English:

If the English sentence 'We are brains in a vat' is true (i.e., our sentence 'We are brains in a vat'), then we speak not English, but vat-English. Since in vat-English 'brains' and 'vat' do not refer to brains and vats, but at best, perhaps, to elements in the machine or image generated by the machine, in vat-English 'We are brains in a vat' is false. That is to say, if our sentence 'We are brains in a vat' is true, then it is, after all, false. Therefore, our sentence 'We are brains in a vat' is false, and necessarily so.

The trouble is that when we attempt to say that if our sentence 'we are brains in a vat' is true, then *it* is false, we are trying to say something at once in one and two languages. A sentence is true or false only as interpreted or relativized to some language. When we talk about our sentence, we talk about a sentence we use interpreted relative to our language. We are trying here to say that if a certain sentence in our language is true, then it is false, where the second reference to the sentence presupposes a shift in the language to which the sentence is relativized from the language in which the antecedent is expressed, and yet at the same time presupposes that it is the same language, the language of the speaker, as that in which the antecedent is expressed. But that is incoherent.²⁵

²⁵ It might be thought that we get out of this problem by making the conditional subjunctive--if we were brains in a vat, then our sentence 'we are brains in a vat' would be false. (i) To the extent that the use of the subjunctive mood implies that we aren't presently brains in a vat, we cannot use this conditional in an argument to show we aren't, since we want to start from premises that don't already commit us to that. (ii) Waiving this, if we take the antecedent to refer to the actual circumstances, then if it is true, our sentence 'we are brains in a vat' is also true, since the conditional is expressed using our language, and consequently the conditional is false. If we take it to refer to non-actual circumstances, then we know something about the truth value of a sentence like our sentence 'we are brains in a vat' in those circumstances, but not in the actual circumstances. Still, the subjunctive conditional is germane to the argument, because the fact, if it is one, on which its truth depends provides premises from which we can deduce we are not brains in a vat.

It looks as if we can make the move that Putnam does because his theory apparently shows that if someone is what we call a brain in a vat, then his sentence ‘we are brains in a vat’ is false, since in vat-English that sentence is false. This applies to everyone, apparently, so it applies to us. So if we are brains in a vat, our sentence ‘we are brains in a vat’ is false because it is in vat-English. The difficulty arises because to state the generalization we have to assume that we are speaking English, not vat-English. There is not the same difficulty in applying it to someone else. We can’t apply it to ourselves because that would mean undermining the assumption that we are speaking English (our language), upon which the coherence our assertion depends.²⁶

But, as I have indicated, there is another way of arguing to Putnam’s conclusion from his theory, bruited in the discussion of Brueckner’s objection, that doesn’t encounter this difficulty. The reason we can’t be brains in a vat is that brains in a vat, as we see from Putnam’s theory, can’t express what we mean by ‘brains in a vat’. To put this another way, since brains in a vat can’t think that they are brains in a vat, but we can, we are not brains in a vat. This follows from the theory because it is a condition on our referring to brains and vats that we not be brains in a vat of the type in Putnam’s story. This way of putting the argument shows that it relies on the claim that the supposition that we are brains in a vat is in a way self-refuting—which is not equivalent to the final form in which Putnam puts his argument. It is a condition on our saying or thinking that we are brains in a vat that we are not brains in a vat. Since we know we can think that we are brains in a vat, we know that we are not.

²⁶ Perhaps there is another way of putting the argument along these lines, taking advantage of the fact that the antecedent in the conditional in English entails that I do not speak English, that is, the language in which I express it. We might put it this way. (1) If I am a brain in a vat, then I do not speak the language in which I am expressing (1). (2) For any proposition P which I express using a sentence S, it is impossible that I fail to speak the language in which I express P with S. From (1) and (2), I am not a brain in a vat. The key point is that a fact that I know to be true is incompatible with my being a brain in a vat. In this argument, I *presuppose* I know what my words mean, and what I think, and rely on the fact that the brain in a vat would not be speaking my language. In the argument in the main text, the focus is on the brain in the vat meaning something different than I do by ‘I am a brain in a vat’. It is crucial to both arguments that I know the meanings of my words. For this reason the argument in the text seems to me to be a more straightforward representation of the basis of the anti-skeptical argument.

7. Why the argument is not successful

I claimed earlier that although Brueckner's objection to Putnam fails, Putnam's theory in conjunction with other assumptions he makes does entail that we don't know the meanings of our words, and that if Putnam gives up those assumptions, his argument for his theory of reference is undercut. This conclusion is not based on the claim that we don't know whether we are speaking English or vat-English if Putnam's theory is correct, but on Putnam's assumption that it is possible that there is some brain in the vat such that it and I have *introspectibly* the same experiences and thoughts, though their representational content differs. Assume, with Putnam, that, although there is no introspectible difference between my mental life and that of some brain in the vat, our thoughts have very different contents. For instance, when we both think 'There's a tree', one of us means that there's a tree, and the other means that there's program feature alpha (or at least something that picks out program feature alpha). Now I do not learn what I think by checking to see what correlations there are between my words and my environment. If there is anything that can be said to be my basis²⁷ for knowing that I have a certain thought content, it is its introspectible quality (or, as I will sometimes say, its subjective character). Indeed, I think what gives content to the idea that there are two different perspectives on a person's thoughts, something we have seen that Putnam needs, is the fact that one has conscious access or knowledge of one's own thoughts while no one else does or can. But, by hypothesis, the introspectible quality of my thoughts is the same as that of one of the brains in the vat. If we think of the introspectible quality of the thought as being my basis for knowing its content, then my basis is not sufficient for distinguishing between the thought that there's a tree and the thought that

²⁷ I mean 'basis' to be given a thin reading. I am not suggesting, nor do I think it intelligible, that there is a character my conscious mental states have from which in some way I infer their representational contents. Any process of inference requires knowledge of content prior to that inferred. Knowledge of content can't, at least in general, then, be inferred from something else. By 'basis' I mean something that can give content to the idea that I have a special access or way of knowing my own thoughts that no one else does or could have.

there's program feature alpha. Consequently, if Putnam is right, I do not know whether my thought 'There's a tree' is about trees (really) or program feature alpha.

To put the argument a little more generally, if we have knowledge of thought contents, then the subjective character of our conscious thoughts determines that knowledge. Knowledge of thought contents determines what the thought contents are. By transitivity, subjective character determines thought contents. So if we have knowledge of thought contents, they are determined by subjective character. Since Putnam's theory denies that thought contents are determined by subjective character, if his theory is true, we do not have knowledge of thought contents.

If we accept that the subjective character of our conscious mental states determines our knowledge of our thoughts and our knowledge of the meanings of our words, then, since it would be absurd to suppose we don't know what we mean and think, we must reject Putnam's theory insofar as it divorces the determinants of meaning from the subjective character of our conscious mental states. One option open to Putnam is to give up the assumption that subjective character remains invariant while content changes. But giving up the assumption undermines the argument Putnam gives for his theory of reference.

To see this, recall the structure of the thought experiments which Putnam describes to support his theory. In the first of these, we were to imagine that someone who had never been in causal contact with trees nonetheless somehow acquired a mental image of a tree. We judge that he does not refer to trees. Since he has the image of a tree, it can't be in virtue of the image of a tree that he refers to trees. He fulfills that condition, but still does not refer to trees. The same thing goes for mental words, and in fact all introspectible mental states and events. We can describe situations in which an individual has all the same mental states and events as our own, as far as their introspectible character goes, but is not thinking the same thing we are. So it is nothing introspectible in virtue of which we refer to or represent things in the world.

It is clear that we only get the conclusion that there is nothing introspectible that determines what someone refers to if we can describe a situation in which what's introspectible remains the same while what a person refers to differs. We can do this without difficulty in the case of reference to

particulars. If I wake up tomorrow morning, having forgotten the previous day, and think to myself, ‘Today is Monday’, I may be in the same introspectible state that I was the day before, when I thought, as we would say, the same thing. But to get Putnam’s conclusions about the brains in the vat, we need also to describe a case in which the extensions of predicates can differ globally when the introspectible state remains the same. I have suggested that in such a case we could not know the extensions of our terms. But in the absence of such a case, we cannot conclude that nothing introspectible determines reference—at least not from the sort of thought experiment Putnam describes.

If introspection affords us knowledge of the contents of our mental states and events, and what we introspect, as Putnam assumes, is the qualitative or subjective character of our thoughts, then the representational content of our thoughts cannot be separated from their subjective character. To put this another way, thoughts and images are individuated in part in terms of their representational content, and it is that content we know through introspection, as we have (in general) no other access to it. It is clear that if introspectible images and thoughts are individuated partly in terms of their representational content, Putnam’s argument won’t go through. For then we cannot describe a case in which two people have the same introspectible thought but different thought contents.²⁸

²⁸ Earl Conee, in a review of *Reason, Truth and History*, in *Nous*, (March 1987) p. 83, makes a related criticism of Putnam. Conee represents the trouble for Putnam as his right to say we can conceive of, say, trees, given what he says about the deliverances of introspection remaining invariant with brain states. Conee seems to be assuming that Putnam has committed himself explicitly to the view that knowledge of the meanings of our words or the contents of our thoughts depends on what is introspectible about our conscious mental states. Nothing Putnam says commits him explicitly to that view. In my criticism of Putnam, I represent this as an independent assumption, one which I believe to be true, and given which, it would be impossible, not only to know the contents of our thoughts if the conclusion of Putnam’s argument were true, but to state coherently Putnam’s argument for his theory of reference. It is, furthermore, an assumption which I see as essential to Putnam in a defense against the *general charge* that his causal theory of reference undermines knowledge of meanings, and therefore thought content.

8. Another look at the thought experiments

If we start with this idea and take another look at Putnam's thought experiments, one thing that should strike us immediately is that the only way Putnam has of identifying the image that is supposed to be qualitatively identical to one he might have of a tree is by saying it is an image *of a tree*. The same thing is true when we consider the description of the thought words that are supposed to go through a person's head which are identical with the ones that go through ours though they mean different things. We identify the thought by saying, e.g., it's the thought 'there's a tree there', where we have a grasp on what thought that is because it is *our* sentence.

Why would Putnam have supposed that one could identify a mental image or thought introspectively apart from its representational content? Putnam starts out by talking about how word inscriptions and pictures represent. Word inscriptions and pictures don't represent intrinsically. The same word inscription or image can represent very different things or even nothing at all, for physical objects have a character that is independent of what, if anything, they represent. This picture motivates Putnam's remarks about the mental. In effect, he assimilates introspectible mental images and thoughts to physical pictures and sentence inscriptions, and assumes that just as physical pictures and sentence inscriptions have a character independently of what they represent, so do mental images and thoughts. Putnam's argument appeals, I think, to an old picture of the mind, the mind as a theater of sorts in which consciousness eyes its inner objects. I am sure that Putnam would want to repudiate this picture. Yet it is this picture that suggests the assimilation of these inner objects to physical objects, and suggests that whatever we say about physical objects, in respect of their representational properties, we should say as well about these inner objects. I think we have only to expose the picture to begin to doubt its propriety. There is no reason initially to suppose that introspectible mental events and states, from which physical objects derive whatever representational properties they have, should, like those objects on which they bestow representational properties, have a character that is specifiable, from the point of view of introspection, completely independently of whatever they represent.

This does not show that mental images and thoughts which we introspect represent or refer to things intrinsically.²⁹ There is the other half of Putnam's argument, that to suppose they do represent or refer intrinsically is to subscribe to an unscientific, magical, and therefore bad theory of reference or representation. This second half of Putnam's argument I do not want to address here. The question it raises sharply is whether causal theories of reference are any less magical than theories that ascribe intrinsic powers of representation to mental states.

There is one further point I want to touch on briefly in the way Putnam sets up his argument. In a footnote at the beginning of the first chapter of *Reason, Truth and History*, Putnam says that he is restricting his notion of 'reference' and 'representation' to something like denotation. It is a commonplace that not all representation requires that we successfully refer to something. 'Whoever is bald is virile' has representational content in this sense, and would, even if no one were bald or virile, and no one ever had been or will be. Likewise, there is a sense in which I mean or represent the same thing you do when I say 'The bald man in the corner is a spy', even if on the occasion on which I say it there is a bald man in the corner, and on the occasion on which you say it he has moved and there is no one in the corner.

Noting the restriction Putnam places on his use of 'reference' and 'representation' encourages a deflationary reading of his argument: all Putnam has shown, it might be said, is that the brain in the vat can't refer to anything unless it is in appropriate causal contact with it. There is a sense in which when the brain in the vat thinks 'There's a tree' it's not thinking about a tree. But this is the same sense in which when I say 'The bald man in the corner is a spy', and there is no one in the corner, I'm not thinking about anyone. But I am still thinking *something*, even if, to avoid a misleading suggestion, we hesitate to say what I think, though clearly not true, is straightforwardly false. The content of my thought is the same as it would have been if there had been a bald man in the corner. Likewise, the brain in the vat is thinking 'There's a tree' and means what we do by that, and fails to be, in one sense,

²⁹ It is important here that we think of this from the point of view of introspection: for this allows *prima facie* (I do not say there are not deeper difficulties) that this vantage on our mental states may not specify them independently of their representational content while allowing that that content is fixed by relational properties.

thinking about a tree simply because his assumption that he is in perceptual contact with a tree is false. The brain in the vat, then, far from having mostly true beliefs in virtue of not being in causal contact with trees and tables and chairs and the like, and so not thinking about such things, has mostly false beliefs precisely because he fails, when thinking about things about him, to be thinking about trees and tables and chairs and so on.³⁰

I think it is clear that Putnam doesn't mean his argument to be so limited. This comes out in his assumption that the brain in the vat would be referring to things after all, which would not make sense on this interpretation of the argument. When laying out Putnam's argument, I assumed on Putnam's behalf at one point that it is a condition on representation of objects in this second sense that we be or have been in causal contact with some of the things we represent, or at least with things in terms of which we could build up the representations. The importance of this is that the assumption is not directly supported by Putnam's thought experiments, which are couched in terms of what a person can refer to, where he can refer to something only if it exists. The persuasiveness of Putnam's thought experiments, I believe, relies on thinking of 'referring' as reference to particulars, which requires that the thing referred to exist. The conclusion drawn employs a broader notion of reference.

9. Conclusion

Putnam's theory of reference and representation has the right form for a response to the traditional problem of skepticism about the external world. It attempts to show that there is a logical connection between the nature of our thoughts and the nature of the world around us. In contrast to traditional accounts of the mind-world relation that secure this connection by making the world depend upon the mind, Putnam's account makes the mind depend upon the world.

I have argued that if Putnam's theory of reference were correct, it would show that we cannot be brains in a vat (of a certain sort), and also, more importantly, that most of our beliefs are true. I

³⁰ Alternatively, one might say he has beliefs that are neither true nor false. But this is obviously no response to skepticism.

have disagreed with a common charge leveled against this kind of theory, that any theory of this kind will have the result that we don't know what we mean by our words or what we think. This charge depends upon the claim that if Putnam's theory is correct, to know what we mean and therefore think we would have to know already what our environment is like, what sorts of things we have been in regular causal contact with. But this incoherent requirement is not entailed by Putnam's theory. The objection can be met by allowing that one knows one's own thoughts and therefore meanings in a way different from the way anyone else does. I cannot see that there is any incompatibility between this claim and the view that one's meanings and thought contents are determined by one's causal relations with one's environment.

However, what grounds the distinction between first and third person knowledge of one's mental states is that one's own mental states are for oneself accessible to consciousness--that is, they can be manifested in consciousness, or as conscious mental states--while they are not and cannot be so accessible to anyone else. This consciousness is not to be distinguished from awareness of a state's subjective character. It follows that to the extent we have first person knowledge of our own conscious mental states, that knowledge is determined by those states' subjectivity. To the extent that the determinants of thought content are divorced from the determinants of subjectivity, we cannot have knowledge of thought content. It is incoherent to suppose we do not know what we mean or think. Therefore, any relation theory of meaning or thought content must hold that subjectivity is relationally determined as well. However, it is an assumption of Putnam's argument that representational content can vary independently of the subjective character of a person's conscious mental states. Consequently, if his argument were correct, we would be unable to know what our thoughts were about. But if Putnam drops that assumption, the thought experiments on which his theory rests are undermined.

The mistake can be traced to Putnam's assimilation of the mental to the physical. We cannot conclude from the fact that physical states or objects can be characterized independently of whatever representational content they may have that the subjective character of mental states can be characterized independently of whatever representational content they may have. I have suggested that

this assimilation rests on an old theory of the mind, one according to which the eye of consciousness is entertained by various mental objects that pass across a mental stage. Putnam of course entertains nothing so crude. But the underlying idea, the underlying tendency, to see mental images, and conscious thoughts, as objects of a sort, conceivable independently of their meaning, just like physical objects, is the same. Once we have rejected this picture of the mind, there is, I think, no further temptation to think that representational content comes apart from subjective character.