CANADIAN JOURNAL OF PHILOSOPHY
Volume 31, Number 4, December 2001, pp. 455-480

Semantic Pragmatism and A Priori *Knowledge*
(or 'Yes we could all be brains in a vat')[1]

HENRY JACKMAN
York University
Toronto, ON
Canada M3J 1P3

I Introduction

Hilary Putnam has famously argued that we can know that we are not brains in a vat because the hypothesis that we are is self-refuting.[2] While Putnam's argument has generated interest primarily as a novel response to skepticism, he originally introduced his brain in a vat scenario to help illustrate a point about the 'mind/world relationship.'[3] In particular, he intended it to be part of an argument against the coherence of metaphysical realism, and thus to be part of a defense of his conception of truth as idealized rational acceptability. Putnam's discussion has already inspired a substantial body of criticism, but it will be argued here that these criticisms fail to capture the central problem with his argument. Indeed, it will be shown that, rather than simply following from his semantic

1 I'd like to thank Robert Brandom, Joe Camp, Jonathan Cohen, Brian Garrett, Mark McCullagh, John McDowell, Ram Neta, Deborah Smith, audiences at the 1999 Mid-South Philosophy conference and York University, and two anonymous referees for comments on earlier versions of this paper.

2 H. Putnam, *Reason, Truth and History* (New York: Cambridge University Press 1981), ch. 1

3 Ibid., 6. This point is stressed in D. Davies, 'Putnam's Brain-Teaser,' *Canadian Journal of Philosophy* **25** (1995), 224-7, and in M. Hymers, *Philosophy and Its Epistemic Neuroses* (Boulder, CO: Westview 2000), ch. 1.

externalism, Putnam's conclusions about the self-refuting character of the brain in a vat hypothesis are actually out of line with central and plausible aspects of his own account of the relationship between our minds and the world. Reflections on intentionality and semantics ultimately give us no compelling reason to suppose that the beliefs expressed by claims like 'I am a brain in a vat' could not be true,[4] but (*pace* Putnam) this supports neither skepticism nor metaphysical realism.

II Putnam's Argument

Putnam's attempt to show that we could not be brains in a vat begins with his asking us to imagine the following scenario:

> [A] human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal.[5]

While this scenario 'violates no physical law,' and is 'perfectly consistent with everything we have experienced,' Putnam still insists that it 'cannot possibly be true, because it is, in a certain way, self-refuting' (ibid., 7). Putnam takes the hypothesis to be self-refuting because it purports to state a possibility that (according to his understanding of semantic externalism) should be unstateable. What our words refer to is determined by what their usage is causally connected to, and a brain in a vat's usage of 'vat' would not have the sorts of causal connections to vats needed for it to designate them. As Putnam puts it, '"Vat" refers to vats in the image in vat-English, or something related (electronic impulses or program features), but certainly not to real vats, since the use of "vat" in vat-English has no causal connection to real vats.'[6] If we were brains in a vat, then our word 'vat' wouldn't refer to *vats*. Consequently, the

---

4 To simplify the presentation, the paper will typically focus on the types of claims that we can make. Nevertheless, the point should be understood as extending to the thoughts expressed by those claims. The relevant issues deal with the limitations on what we can think about as much as they do with what we can coherently talk about.

5 *Reason, Truth and History*, 5-6

6 Ibid., 14. See also, 'Although the people in that possible world can think and "say" any words we can think and say, they cannot (I claim) *refer* to what we can refer to. In particular, they cannot think or say that they are brains in a vat (*even by thinking "we are brains in a vat"*)' (ibid., 8).

mere fact that we can raise the possibility that we are brains in a vat shows that we are not. In other words, 'If we can consider whether it is true or false, then it is not true.... Hence it is not true.'[7]

Putnam's assumption that the use of 'vat' in vat-English has no causal connection to real vats is, of course, essential to his argument, and this leads him to seriously modify his original scenario. Someone 'subjected to an operation by an evil scientist' could have had plenty of causal contact with vats before being envatted, and even someone who was always a brain in a vat could have such causal connections indirectly (through artificial sense receptors or changes in the virtual environment based upon what takes place outside of it). It is, then, not surprising that Putnam adds a number of further embellishments to his story. In particular, one's brain is supposed *always* to have been envatted, and the vat and automated machinery are no longer designed by an intelligent scientist, but rather are 'supposed to have come into existence by some kind of cosmic coincidence' so that they 'have no intelligent creators or designers' (ibid., 12). It is only the hypothesis so embellished that is supposed to be self-refuting. Consequently, in spite of Putnam's tendency to say things such as 'I am claiming that there is an argument we can give that shows that we are not brains in a vat' (ibid., 8), Putnam presents no such argument unless 'brains in a vat' is understood as shorthand for the modified scenario.[8] Nevertheless, the secondary literature has followed Putnam's lead in using the expression 'brain in a vat' to refer to the modified scenario, and this paper will, henceforth, do the same.

III Traditional Objections and the Shared Assumption

Even those who typically see little point in worrying about whether or not we might all be brains in a vat have been surprised by Putnam's claim to be able to *prove* (indeed, prove *a priori*) that we couldn't be. Conse-

---

7 Ibid. See also: 'It follows that if ... we are really 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.... So, if we are brains in a vat, 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' (ibid., 15).

8 For a discussion of this, see A. Brueckner, 'Brains in a Vat,' *Journal of Philosophy* **84** (1986), 152.

quently, many have attempted to reconstruct his argument more formally in order to identify precisely what assumptions and inferences it requires. There have been many such reconstructions, but Putnam's argument can, for present purposes, be reformulated as follows:[9]

(i) My language disquotes.

(ii) In vat-English, 'brain in a vat' does not refer to brains in vats.

(iii) In my language 'brains in a vat' is a meaningful expression.

(iv) In my language, 'brains in a vat' refers to brains in a vat. [From (i) and (iii).]

(v) My language is not vat-English. [From (ii) and (iv).]

(vi) If I am a brain in a vat, my language, if any, is vat-English. [Definition of vat-English.]

(vii) I am not a brain in a vat. [From (v) and (vi).]

While this argument seems valid, questions about the types of semantic self-knowledge compatible with the externalist semantic framework Putnam presupposes have led some to challenge our *a priori* entitlement to a number of its steps.

Possibly the earliest and most influential line of objection to Putnam's argument questions our *a priori* entitlement to (iv).[10] We do not have introspective access to those 'external' factors that, according to Putnam's externalistic semantic framework, determine what our expression 'brain in a vat' refers to. Consequently, even if Putnam's argument lets one know that one's sentence 'I am not a brain in a vat' must be true, it doesn't let one know that one is not a brain in a vat. One may know the sentence's truth-value, but one still lacks *a priori* access to its *content*. Disquotation alone is not enough to insure that one knows what one's sentences mean, since mastery of the disquotation schema does not require understanding all of the terms found within it. To *really* know

---

9 I am here following C. Wright, 'On Putnam's Proof that We Are Not Brains in a Vat,' P. Clark and B. Hale, eds., *Reading Putnam* (Cambridge: Blackwell 1994), 224, since Putnam himself seems to endorse this reconstruction (H. Putnam, 'Comments and Replies,' Clark and Hale, eds., 284). In any case, the objections considered below should be locatable in any of the many acceptable formulations of Putnam's argument available.

10 For the best known exposition of this line, see Brueckner, 'Brains in a Vat.'

that one was not a brain in a vat, one would have to know whether one was speaking English or vat-English, and one could only know that if one already knew whether or not one was a brain in a vat.[11]

A second, and more radical, line of attack can be directed at step (iii). Critics can argue that our lack of semantic self-knowledge extends to the point that one cannot even tell by introspection whether one's words and 'thoughts' are contentful at all. For all one knows, one's words and thoughts may have none of the causal connections to the external world needed to make them meaningful.[12] How, the critic of our entitlement to (iii) might ask, does one know that one is not just 'thinking' with words that are utterly lacking in content? Internalists could at least be sure that they were thinking, but within Putnam's semantic framework, one's claim to know *a priori* that the expressions running through one's consciousness are meaningful, and thus one's entitlement to step (iii), seems undermined.

The force of such attacks on Putnam's argument is a matter of some controversy.[13] However, even if such criticisms are sound, they still allow Putnam's argument to establish a number of surprising and non-trivial conclusions. The first objection allows that any attempt to *formu-*

---

11 'I can conclude ... that I am a normal human being rather than a BIV ... only if I can assume that I mean by "I may be a BIV" what normal human beings mean by it. But I am entitled to that assumption only if I am entitled to assume that I am a normal human being speaking English rather than a BIV speaking vat-English. This must be shown by an anti-skeptical argument, not assumed in advance' (Brueckner, 'Brains in a Vat,' 103).

12 The connection between externalism and this possibility is made very vivid in the discussion of the 'Swampman' in D. Davidson, 'Knowing One's Own Mind,' P. Ludlow and N. Martin, eds., *Externalism and Self-Knowledge* (Stanford: CLSI Publications 1998). Putnam discusses a related possibility in *Reason, Truth and History*, 17. The possibility of externalist anti-skeptical arguments backfiring in this way is discussed by Brueckner, 'Brains in a Vat,' 159; Falvey and Owens, 'Externalism, Self-Knowledge, and Skepticism,' *Philosophical Review* **103** (1994) 107-37, at 126; P. Klein, 'Radical Interpretation and Global Skepticism,' LePore, ed., *Truth and Interpretation* (Oxford: Blackwell 1986), 385; and S. Stich, 'Might Man Be an Irrational Animal?' H. Kornblith, ed., *Naturalizing Epistemology* (Cambridge: The MIT Press 1994), 356.

13 See, for instance, Brueckner, 'Brains in a Vat'; T. Nagel, *The View from Nowhere* (New York: Oxford University Press 1986); Falvey and Owens; Wright, 'On Putnam's Proof'; Davies, 'Putnam's Brain-Teaser'; G. Forbes, 'Realism and Skepticism: Brains in a Vat Revisited,' *Journal of Philosophy* **92** (1995); G. Ebbs, *Rule Following and Realism* (Cambridge: Harvard University Press 1997); H. Noonan, 'Reflections on Putnam, Wright, and Brains in Vats,' *Analysis* **58** (1998); S. Sawyer, 'My Language Disquotes,' *Analysis* **59** (1999); Ludlow and Martin; and the numerous papers cited therein.

late the skeptical hypothesis will be false. The second allows that a denial of the skeptical hypothesis is presupposed by our assumption that we are thinking at all. Both objections thus leave in place the conclusion that there is something fundamentally problematic with attempts to *claim* that one might be a brain in a vat. Each concedes that if one is entitled to the claim that one is thinking (and that one knows *what* one is thinking), then one is entitled to the claim that one is not a brain in a vat. Such concessions are substantial (indeed, they are too substantial), and the problems with Putnam's argument are more fundamental than these two standard objections suggest.

In particular, the most serious problem with Putnam's argument is with step (ii), namely the assumption that:

(ii) In vat-English, 'brain in a vat' does not refer to brains in vats.

Putnam, his supporters, and his critics typically agree that one can know *a priori* that, if one were a brain in a vat, then one's word 'vat' would not refer to vats. Consequently, they all assume that the claim 'I am a brain in a vat' couldn't *possibly* be true. Their disagreements are over what this purported 'semantic' fact is supposed to show. Putnam and his sympathizers take it to show that we can know that we are not brains in a vat, while his critics take it to show only that we can know that a certain type of utterance, if meaningful, must be false.

It is this shared assumption, that a brain in a vat could not refer to brains in vats, and thus could not truly think 'I am a brain in a vat,' that should be questioned. Indeed, the problems with (ii) are considerably more serious than those with (iii) and (iv). I know of no one who seriously questions the *truth* of (iii) or (iv). All that is questioned is our *a priori* entitlement to them. I can doubt (iii) and (iv) only in some very limited 'philosophic' sense. I recognize that I may not be entitled to them while in a philosophical argument with the skeptic, but I have no *real* doubt about their truth. No one really doubts that the word 'vat' is a meaningful expression, or that it refers to vats. On the other hand, many people's naïve intuitions about (ii) seem to be that it is false. This is why the Evil Demon and Brain in a Vat hypotheses have seemed coherent, if implausible, to so many. Such naïve intuitions can, of course, turn out to be incorrect. Nevertheless, it will be argued here that not only do we have little reason to think that (ii) can be known *a priori*, but we also have good reason to doubt that it is true at all. The following criticisms of (ii) (unlike those of (iii) and (iv)) thus question not only the *a priori* availability of Putnam's argument, but also its *soundness*.

The hypothesis that I am a brain in a vat *seems* like an intelligible one, and the *prima facie* intelligibility of the hypothesis partially explains the intuitive discomfort that many people have with Putnam's purported

proof of its incoherence.[14] Indeed, given the rather unpromising history of attempts to rule out such 'skeptical' scenarios on semantic grounds,[15] one might think that any account of meaning that entailed that the brain in a vat hypothesis was unintelligible would, thereby, cast serious doubts upon its own acceptability.[16] Questioning (iii) and (iv) does not, however, get at what is intuitively suspect about Putnam's argument, since there is nothing unintuitive about Putnam's assumptions that our words are meaningful and that *our* word 'vat' refers to vats. Rather, what is unintuitive is the claim that a brain in a vat would be saying something false were it to say 'I am a brain in a vat.' The objections that focus on (iii) and (iv) typically endorse (or at least ignore) this claim, and only question what sort of knowledge can be derived from it. By contrast, an attack on (ii) gets to the heart of the matter by defending the conceivability of the brain in a vat's ability to make a true claim about its condition.

IV Switching and De-vatting

Before evaluating the plausibility of premise (ii), consider the following two cases.

1. A speaker discovers that seven days ago, while sleeping, he was transported (it doesn't matter how) to Earth from his own planet (hereafter 'Earth2'). Earth2 seems just like Earth though every substance on it has a different atomic structure than does its Earth-counterpart. On looking back at what he said over the past week, he is inclined to say that assertions like 'I'm in Toronto,' 'That's Hilary Putnam,' 'Here is a rabbit,' and 'This is water' were mistaken. He considers his terms 'Toronto,' 'Hilary Putnam,' 'rabbit,' and 'water' to not (yet) refer to the people,

---

14 After all, most people (including Putnam himself: *Reason, Truth and History*, 7) typically *do* feel that there must be something ultimately wrong with Putnam's argument when it is first presented to them. (Or at least that is my experience with students when they are presented with the argument, and with most of my colleagues with whom I have discussed it.) Their intuition is that there must be *something* wrong with the argument, even if they cannot pin down precisely what that something might be.

15 For a discussion of some of these, see B. Stroud, *The Significance of Philosophical Skepticism* (New York: Oxford University Press 1984).

16 See Falvey and Owens; C. McGinn, 'Radical Interpretation and Epistemology,' LePore, ed.; and M. Williams, *Unnatural Doubts* (Cambridge: Blackwell 1991), xiv, for claims of this sort. However, I will ultimately argue that a commitment to 'semantic externalism' is perfectly compatible with the hypothesis's intelligibility, and thus that no *reductio* of semantic externalism is in the offing.

places, animals, and substances that go by those names here on Earth. Rather, he takes them to refer to their counterparts on Earth2. Nevertheless, he thinks that, say, the things he called 'phones,' 'cars,' and 'spoons' here on Earth were, in fact, phones, cars, and spoons. Indeed, he thinks that he was correct to call the vats on Earth 'vats,' even if he could not truly apply any of the terms for what Earth2-vats are made of ('copper,' 'steel,' 'iron,' etc.) to the vats on Earth.

2. A speaker discovers that seven days ago (it doesn't matter how) his sleeping brain was scooped out of the vat it had always floated in and placed in a human body. He discovers that while his new environment seems exactly like his old one, his experiences of his old environment were dependent upon the states of a computer in this new environment. Indeed, the whole set-up responsible for his previous experience seems to have come together through some sort of 'cosmic coincidence'. Looking back at what he said over the past week, he is inclined to say that assertions like 'I'm in Toronto,' 'That's Hilary Putnam,' 'Here is a rabbit,' and 'This is water' were mistaken. He takes his terms 'Toronto,' 'Hilary Putnam,' 'rabbit,' and 'water' not to refer to the people, places, animals, and substances in this environment. Rather, he takes them to refer to their counterparts in his previous computer-generated environment. Nevertheless, he still thinks that could correctly identify the phones, cars, spoons, and vats in this new environment as 'phones,' 'cars,' 'spoons,' and 'vats.'[17]

These two cases may represent how speakers would describe themselves and their usage upon discovering that they had been recently 'switched' or 'de-vatted.' The question remains, however, of whether we should endorse such descriptions.

V Externalism and Non-Natural Kinds

Of course, the intuition that someone could correctly identify vats as 'vats' upon being 'de-vatted' is precisely what Putnam claims semantic externalism gives us compelling reasons to reject. However, the intuition can be understood as compatible with semantic externalism if we understand 'vat' to pick out some type of kind that has instances in both 'real' and 'virtual' contexts. Indeed, it will be argued below that while a natural

17 I should note that this case departs from Putnam's example slightly since it allows that there are other conscious creatures outside of the vat. Nothing, however, should turn on this, since the creatures outside of the vat are taken to have nothing to do with the vat in which the brain sits and the virtual world generated by the computer.

kind term like 'water' may be inapplicable in contexts where the functionally/experientially equivalent substances lack water's molecular structure, terms like 'spoon' or 'vat' may be applicable 'across contexts' provided that the differently constituted 'spoons' or 'vats' play a relevantly similar 'role' in the alternate environments.

The sampling of objects that one's usage is causally dependent upon constrains what one's terms can refer to, but it does not, in itself, determine what sortals they should be interpreted as falling under.[18] A term in a language can denote objects that its users have not had causal contact with if it picks out a category/kind that encompasses both those unexperienced objects and whatever instances of the kind that the speakers have experienced. There are, after all, numerous types of kind that our terms could pick out, and I will here mention just three. First of all, a term could pick out a 'natural kind' of the sort determined by the microstructure of the initial sample.[19] On the other hand, it could pick out a 'functional kind' that was, say, sensitive to aspects of the causal role played by members of the initial sample. Finally, it could pick out an 'interactional kind' that picked out objects that interacted with speakers in ways relevantly similar to the initial sample. (Functional and interactional kinds are in many ways quite similar, though important differences between the two will be explained later.) The objects that a term has actually been applied to can often be understood as instances of all three kinds of kind. As a result, the question of what kind of kind a term picks out often comes down to the question of what aspects of the initial sample do the user(s) of the term find the most important when they are applying it, and thus which unexperienced objects would they find relevantly similar to the initial sample. We would not take 'water' to apply to a functionally similar but molecularly different substance on another planet because the similarity we take to be relevant to the application of 'water' is microstructural similarity. However, there is no reason to think that this type of similarity governs our application of all of our terms.

For instance, while our use of 'vat' has no causal connection to the vats on Earth2, that hardly means that they cannot fall within the term's extension. If our term 'vat' were interpreted as picking out, say, some sort of functional or interactional kind, then 'vat' would pick out both the vats on Earth that we have experienced and the vats on Earth2 that

---

18 A point that should be familiar from Putnam's own discussions of the feasibility of *purely* causal accounts of reference. (See, for instance, *Reason, Truth and History*, 53.)

19 I doubt that terms like 'natural kind' or 'functional kind' themselves pick out 'semantically natural kinds,' and so these suggestions are not meant to be capture of how all such kind terms should be characterized.

we have not. In much the same way, the vat-English term 'vat' might pick out a kind that includes both the vats-in-the-image (or 'virtual vats')[20] that the speakers of vat-English have experienced, and the 'real vats' that they have not.[21] All that is needed is for the speakers of Vat English to be disposed to consider (on reflection) the non-virtual vats to be relevantly similar to the virtual vats that they are used to.

After all, consider the following three sets of objects: the set of all physical vats (hereafter P-Vats), the set of all virtual vats (hereafter V-Vats), and the combined set of all P- and V-Vats (hereafter C-Vats).[22] While the brain in a vat's term 'vat' could not be interpreted as picking out just the set of P-Vats, it is far less clear that it must be interpreted as picking out the set of V-Vats rather than C-Vats.[23] Since all V-Vats are C-Vats, the brain in a vat's usage has had just as much causal contact with C-Vats as it has with V-Vats (while it has had none with P-Vats). In light of this, we should keep in mind that Putnam's premise (ii) is:

---

20 Putnam is less than clear about what these 'virtual vats' should themselves be understood to be. Vats in the image, electronic impulses, and program features have all been suggested by Putnam (*Reason, Truth and History*, 14), Davidson (according to R. Rorty, 'Pragmatism, Davidson and Truth,' *Objectivity, Relativism and Truth* [New York: Cambridge University Press 1991]), and others. I will try not to take a stand on this issue, and will just treat 'virtual vats' to pick out whatever is *causally responsible* for the brain's 'vat utterances.' Consequently, the 'virtual vats' should not be understood as in any way fictional in the way that we think of unicorns as fictional, since, whatever they are, they have causes and effects.

21 One might question this use of 'real vat' and thus Putnam's claim that 'the use of "vat" in vat-English has no causal connection to real vats' (*Reason, Truth and History*, 14). Both may seem to beg the question at hand by assuming that the vats in the image could not be 'real.' On the other hand, one might try to preserve Putnam's claim by arguing that 'real' could be used as a comparative term picking out a *type of* vat, and something that was not a 'real' vat could still 'truly' be a vat. On the various uses of 'real' see J.L. Austin, *Sense and Sensibilia* (Oxford: Oxford University Press 1962), ch. 7.

22 For ease of exposition, assume that we are not brains in vats, and that 'physical' refers to *this* environment, while 'virtual' is virtual *relative to* this environment.

23 Of course, one might try to argue that the categories of P- and V- Vats are somehow more 'natural' than that of C-vats, and that the initial samples only 'project' to such 'natural' properties. See, for instance, D. Lewis, 'New Work for a Theory of Universals,' *Australasian Journal of Philosophy* **61** (1983) and 'Putnam's Paradox,' *Australasian Journal of Philosophy* **62** (1984). However, such a line could hardly be appealed to by Putnam, since such an interest-independent 'ranking' of properties is one of the characteristics of Metaphysical Realism he is most anxious to reject. One of the main themes in *Reason, Truth and History* is precisely that there are no such 'objective' degrees of similarity.

(ii) In vat-English, 'brain in a vat' does not refer to brains in vats.

Which is incompatible with the claim that their term picks out C-Vats. It is *not* the more plausible premise

(ii)* In vat-English, 'brain in a vat' does not refer *exclusively* to P-Vats.

Which is compatible with the expression picking out C-Vats, but not with its picking out just P-vats.[24] Unfortunately for Putnam, while (ii)* is more defensible than (ii), his argument is invalid if (ii)* is substituted for (ii).[25]

In light of this, consider Putnam's analysis of the extension of 'water':

> We can understand the relation $same_L$ (same liquid as) as a cross-world relation by understanding it so that a liquid in W_1 [World 1] which has the same important physical properties (in W_1) that a liquid in W_2 possesses (in W_2) bears the $same_L$ to the latter liquid ... an entity x, in an arbitrary possible world, is *water* if and only if it bears the relation $same_L$ (construed as a cross-world relation) to the stuff *we* call "water" in the *actual* world.[26]

While Putnam may be right to claim that the '$same_L$' relation has to do with physical/micro-structural properties in the case of 'water,' the

---

24 Note that Putnam assumes, in 'Realism and Reason,' *Meaning and the Moral Sciences* (Boston: Routledge & Kegan Paul 1978), 127, that the metaphysical realist would describe the brain in the vat as referring to P-Vats rather than C-Vats by 'vat.'

25 The issue of how to understand premise (ii) is actually more complex that this. One might argue that all that Putnam's argument requires is that the phrase 'brains in a vat' have different extensions in English and Vat-English. (Indeed, Wright suggests something like this in 'On Putnam's Proof,' 221-3.) Consequently, as long as the reference of 'vat' in English was the set of P-Vats, then the argument would go through whether the Vat English expression referred to either C or V-Vats. However, if one takes this line (and I would argue that establishing that the English expression picks out just P-vats is a non-trivial task) the problem reemerges in terms of the question of determining what the referent of 'vat-English' is supposed to be. If 'vat-English' is simply the language spoken by any brains in vats that I would encounter in *my* environment, then the argument is not an interesting one, since there was never a worry about whether I was a brain in one of the vats *in my current environment*. On the other hand, if 'Vat English' is just a general term for English-like languages spoken by anything my expression 'brain in a vat' can truly apply to, it is less clear that it picks out a single language that can be identified with the version of English that would be spoken by any brains in vats found in *this* environment.

26 H. Putnam, 'The Meaning of "Meaning,"' *Mind, Language and Reality* (New York: Cambridge University Press 1975), 232

same-kind relation for 'vat' is not best understood this way.[27] Putnam claims that the 'hidden structures' determine the reference of natural kind terms not because only such hidden structures could serve in the same-kind relation, but rather because 'normally the "important" properties of a liquid or a solid, etc., are the one's that are structurally important.' However, while Putnam's claim may be true for terms like 'water' and 'gold,' importance is, as Putnam himself goes on to stress, 'an interest relative notion,'[28] and for vats it is how we are able to interact with them rather than micro-structural properties that are important. Consequently, we might give the following account of the extension of the brain in a vat's term 'vat':[29]

> We can understand the relation $same_R$ (same role as) as a cross-environmental relation by understanding it so that an object in E_1 [Environment 1] which has the same important interactional properties (in E_1) that an object in E_2 possesses (in E_2) bears the $same_R$ to the latter object ... an entity x, in an arbitrary possible environment, is *a vat* if and only if it bears the relation $same_R$ (construed as a cross-environmental relation) to the things *we actually* call "vats."

How one is able to interact with an object depends upon one's body as well as the object itself. Virtual vats may thus have the same interactional properties as non-virtual vats because the subjects in the vat-world have virtual bodies that interact with them in ways relevantly similar to the ways that non-virtual bodies interact with non-virtual vats.[30] Consequently, if the brain in a vat's term 'vat' picks out this sort of cross-environmental interactional kind, then it would pick out both the 'virtual'

---

27 After all, 'vat' would seem to mean the same thing in English and its Earth2 counterpart, even if the term was applied to an entirely different set of objects on Earth2.

28 Putnam, 'The Meaning of "Meaning,"' 239. This emphasis on our interests separates him, to his credit, from M. Devitt, *Designation* (New York: Columbia University Press 1980) and D. Lewis, 'New Work' and 'Putnam's Paradoxes.'

29 Cross-environmental relations replacing cross-world ones here, since the same object could have different interactional properties in different environments within the same world.

30 The sort of similarity has more to do with the way the subjects involved experience (or would experience) the interaction. (Hence the term 'interactional kind' might also go by 'experiential kind.') As discussed below, the types of interaction might 'objectively' be quite different, since, for instance, what goes on when a non-virtual body kicks a non-virtual vat is significantly different from what goes on when a virtual body 'kicks' a virtual vat. Nevertheless, these two different relations are experienced by the subjects involved in a way that would naturally strike them as relevantly similar.

vats in its own environment and the 'real' vats in ours, because when the subjects change environments, they change bodies as well. If a brain in a vat's terms did pick out such interactional kinds, then, it would be able to truly claim 'I am a brain in a vat.'[31]

Of course, one might suggest that a term like 'vat' need not be analyzed in quite this way. In particular, one could argue that it picks out a functional kind that makes reference not only to the function played, but also to its being able to play it in a *particular* environment.[32] For instance, consider the following account of the extension of 'vat':

> We can understand the relation $same_F$ (same function as) as a environment-specific relation by understanding it so that an object in E_1 [Environment 1] bears the $same_F$ relation to an object in E_2 if it would have the same important functional properties were it in E_2 that the latter object possesses (in E_2) ... an entity x, in an arbitrary possible environment, is *a vat* if and only if it bears the relation $same_F$ (construed as an environment-specific relation) to the things *we actually* call "vats."

If 'vat' picks out this sort of environmentally specific functional kind rather than a cross-environmental one, then 'vat' would have completely different extensions in English and vat-English. Vats would have the function of holding, for instance, *water*, while virtual vats would lack the ability to hold any such non-virtual liquids. The functional roles played would thus be very different. However, it seems unclear why we should believe that by 'vat' the brains in the vat must intend to pick out this more restrictive sort of functional kind rather than the more expansive

---

31 There are, of course, some 'skeptical' hypotheses that *might* still be self-refuting. For instance, the claim 'I have always been a brain in a vat on the dark side of the moon' may be self-refuting. Even if a brain in the vat could talk about our moon as '*a* moon,' when it uses the term 'the moon' it refers to something in the image, not in our world. Much the same could be said of the hypothesis 'I have always been a brain in a vat sitting in *Hilary Putnam's* basement.' Such hypotheses make reference to certain *particulars* in our environment and thus require *names* for their formulation. However, the interest of *these* skeptical hypotheses is obscure to me. Furthermore, there may very well be future experience (discovering massive and constant switching, etc.) that would lead us to conclude that our 'proper names' actually were multi-realizable kind terms.

32 Some seem to think that the term 'functional kind' should only be used this way; hence my preference for 'interactional' kind for the more flexible class of terms. There is, I should note, nothing about the more restrictive account's reference to an actual environment that makes it more in keeping with Putnam's 'indexical' account of meaning than the first. Each type of term allows that the objects experienced in the initial environment help determine the reference, the disagreement is just over which type of sortal these initial samples should be understood in terms of.

sort of interactional kind suggested above.[33] Some functional kinds are 'objective' in that the relevant functions can be specified without making any reference to our activities. For other kinds, the relevant functional role makes essential reference not just to other objects in the world but to how they interact with the experiencing subjects (and possibly their social practices) as well. The resulting kinds may seem very 'unnatural' since, 'objectively,' real and virtual vats (large metal containers and states of a computer) seem to have nothing in common. However, while the interactional kind might seem to pick out 'funny' disjunctions of properties if the experience of the environment-switching subject is left out, that makes them no less legitimate. After all, it has been argued by many (including Putnam himself) that something like this is true of our color terms.[34] Objectively, the things that are, say, blue (my shirt, the sea, the sky) may seem to have little in common, and it is only how they interact with us and our optical apparatus that grounds their falling

---

33 Furthermore, such an understanding of functional kinds would seem to miss out on how we understand even such basic functional kinds such as 'heart.' While human hearts and mouse hearts play similar roles in their respective environments, they could not, to put it mildly, play their roles adequately if their environments were switched. Such an interpretation would also require that the meaning of 'phone' has changed over the last 20 years, since many phones we now use (cell phones in particular) would not be able to function in a remote or past environment where there were no satellites to support them. A less environmentally restrictive account of functional kinds, on the other hand, could easily explains why we are entitled to consider cell phones to be a type of phone. These considerations are hardly conclusive, but as will soon become clear, for the purposes of the current argument, the cross-environmental interpretation of 'vat' need only be established as *possibly* correct to cause problems for Putnam's argument.

34 For a discussion of such cases, see H. Putnam, *The Many Faces of Realism* (LaSalle: Open Court 1987), 5-6; E. Thompson, *Color Vision* (New York: Routledge 1995); G. Lakoff and M. Johnson, *Philosophy in the Flesh* (New York: Basic Books 1999). Furthermore, there is now considerable evidence that classification is often not carried out in terms of categories defined in terms of shared sets of properties; see E. Rosch, 'Family Resemblances: Studies in the Internal Structure of Categories,' *Cognitive Psychology* **7** (1975), and G. Lakoff, *Women, Fire and Dangerous Things* (Chicago: University of Chicago Press 1987). I do not have the space to pursue this point here, but if one accepts such 'prototype driven' accounts of concepts and categories, it would be even easier to defend the claim that a term like 'vat' could be truly applied within the new environment. Much the same could be said of the more 'open textured' account of concepts defended in, for instance, C. Travis, *The Use of Sense* (New York: Oxford University Press 1989) and H. Jackman, 'Semantic Norms and Temporal Externalism,' PhD Diss (University of Pittsburgh 1996) and 'We Live Forwards but Understand Backwards: Linguistic Practices and Future Behavior,' *Pacific Philosophical Quarterly* **80** (1999).

under a single kind. A subject capable of switching environments would just make this phenomenon even more common.

Interactional kinds are easy to conflate with functional kinds since the environment specific causal powers of an object will typically seem to determine its interactional properties. This is because the make up of the other partner in the interaction (our bodies) is usually taken to be fixed. The question of whether a term is a functional or an interactional kind is one that we can typically ignore because of the background assumption that our bodies (and thus the focus of our agency) will not (and could not) change. However, this is precisely the assumption that the brain in a vat scenario calls into question. Just as we can reinterpret what we thought to be a physical kind as a functional one if we discovered it to be multiply realizable, we may decide to treat some of our terms as interactional kinds once we discover that our agency can be focussed through entirely different sorts of 'bodies.'[35]

At this point, one should note that, even if the brain in a vat's term 'vat' could refer to *our* vats, it need not follow that *our* term 'vat' must also refer to the virtual vats.[36] After all, our inclination to understand our term as a cross-environmental one upon discovering a 'virtual world' may be considerably less than the inclination of the recently de-vatted speakers to see their own terms this way. We may ultimately decide that the terms in vat-English typically have different extensions than the terms in English do, but that may only be because the terms in vat-English apply in both environments while the terms in English apply in just one.

Indeed, when traveling between such 'orders of reality,' we may generally be more willing to 'trade up' than 'trade down.' This asymmetry may have to do with the fact that it would be easier for us to view our talk upon entering the vat world as being not quite 'literal' than it

---

35 This may require a fairly stable type of switching between the environments, 'virtual vats' only being encountered when one is 'in' one's 'virtual body,' etc. The stability of such kind terms is thus dependent on some fairly contingent features of our situation. This is, however, arguably a feature of much of our language. For some suggestive discussions on this theme, see L. Wittgenstein, *Philosophical Investigations*, 3rd ed. (Oxford: Blackwell 1953), and J.L. Austin, *Philosophical Papers* (Oxford: Oxford University Press 1961).

36 Though it is far from clear that it shouldn't. After all, if we were to enter into the vat's virtual world, we probably would use regular English words to describe the 'virtual' phones, cars, and vats that we experienced. This raises the question of why we shouldn't simply understand these words in terms of the experienced similarities that lead us to apply them cross-environmentally. (Once again, for ease of exposition, I'm assuming here that we are not brains in vats.)

would be for the 'vat-worlders' to make a similar claim about their talk in ours. This is because it can seem natural to view our talk as 'just pretense' when we enter a virtual environment, but not when we enter an environment relative to which ours is virtual. After all, 'pretend worlds' are not supposed to affect the 'real' ones. A world that is virtual relative to ours has just this quality of pretense. Our own world is causally insulated from that world, and that world could be entirely destroyed without affecting ours. Things are very different from the perspective of the virtual world. Even if the virtual world has traditionally been causally isolated from the goings on of the non-virtual one, it is still causally vulnerable to it. A swift kick or a pulling of a plug in the 'real' world could wipe out the virtual one, and the virtual world could clearly not survive the destruction of the non-virtual one. Because the virtual world is causally embedded in the non-virtual one, it is harder for its inhabitants to treat the objects in the 'non-virtual' world as 'pretend objects' while it is comparatively easy for us to treat our talk of virtual objects as part of an elaborate pretense.

The suggestion that 'vat' in Vat-English could have instances in both environments while our term may only be used correctly (or at least literally) in ours, applies as well for the other terms in the skeptical hypothesis such as 'brain,' 'in,' and 'cause.' (Putnam doesn't treat these terms in much detail, and they will only be discussed briefly here). For instance, Putnam suggests that a brain in the vat would have had no experience of one thing actually causing another, and so could not mean what *we* (purportedly) do by 'cause.'[37] This may be so, but upon being de-vatted, the former brain in a vat might come to realize that what *it* meant by 'cause' was, after all, something (roughly) like law-like correlation. While what the brain in a vat means by 'cause' is not what *we* mean by the term, it could still truly claim that its experiences were being 'caused' by a computer, since its experiences would be correlated with the computer's states in a law-like fashion. Even if it didn't have *our concepts*, it could still use *its concepts* to make claims that were true of *our world*. 'Brain,' too, can plausibly a type of functional/interactional kind term, and brains in a vat might plausibly be able to refer to their actual brain. This is not only because brains are the source of their thinking, which is what they presume the 'brains' to be, but also because, just as one's 'body' could be understood in relation to its interactions as the focus of agency within an environment, one's brain can be understood as whatever plays a certain role vis-a-vis that body. 'In' is also easily

37 H. Putnam, 'Comments and Replies,' 287

understood in experiential terms, and would thus lend itself to a cross-environmental interpretation.

The claim that a brain in a vat could refer to brains and vats thus involves neither a retreat from the 'externalism' upon which Putnam bases his argument, nor the acceptance of any sort of 'magical' or 'transcendental' theory of the mind's relation to the world.[38] 'Interactional' kind terms are still externalistically understood, and the kinds that they pick out are constrained by the *actual* role played by the items referred to in the primary experiential environment. What determines the term's extension is thus not simply the speaker's *conception* of the term. If 'vats' in the virtual environment played a different role than our vats (say, they played the role of industrial colanders whose tiny holes would prevent them from effectively holding liquids), then someone whose experience was limited to those virtual 'vats' would not refer to vats with their term 'vat.' Furthermore, this would hold true even if they were unaware of the tiny holes or straining function of the 'vats' in their own environment.[39] It would still be the vats we *actually experience* that would help determine the interactional kind picked out by 'vat,' not just our *conception* of them.[40]

---

38 For Putnam's use of these terms for any account that would allow the brains in the vat to refer to vats with 'vat,' see *Reason, Truth and History*, 3, 5, 15, 16, and 'Comments and Replies,' 287.

39 If one of them were to say 'we might all be brains in a vat,' another could correctly reply, 'we couldn't be, since all the nutrient fluid would flow out of the holes.'

40 This is why the analysis should not be viewed as 'phenomenalistic' (though a phenomenalistic analysis of the terms in one's language might seem more plausible if one didn't have a single experiential environment). The *cause* of the phenomenon still helps determine the term's extension. Some have argued that non-natural kinds should not be viewed as 'indexical' in this way—S. Schwartz, 'Putnam on Artifacts,' *Philosophical Review* **87** (1978) and 'Natural Kinds and Nominal Kinds,' *Mind* **89** (1980); and M. Devitt and K. Sterelny, *Language and Reality* (Oxford: Blackwell 1987) — and that the extensions of such 'nominal kinds' are 'determined by an analytical specification of superficial features such as phenomenal properties, and/or form, function, or origin' (Schwartz, 'Natural Kinds and Nominal Kinds,' 182). Such accounts, however, must assume that we (or at least some member of our community) can know *a priori* what the relevant functional or formal properties are, and there is no reason to think that this must be (even if it often is) the case. I won't defend this last claim at length here, but the point is developed in H. Kornblith, 'Referring to Artifacts,' *Philosophical Review* **89** (1980), and H. Jackman, 'Semantic Norms.'

VI Semantic Pragmatism and Self-Knowledge

Once again, even if the brain in the vat's term 'vat' would pick out a cross-environmental interactional kind, it need not follow that *our* term 'vat' does so. Nevertheless, there may be no way to tell *a priori* that it doesn't, and if we can't know *a priori* what kind of kind 'vat' is, then we can't know *a priori* that the claim 'I am a brain in a vat' must be false. Indeed, given that some sort of cross-environmental interpretation of 'vat' would seem extremely compelling if we were to suddenly find ourselves de-vatted, the claim that 'vat' is not such a kind presupposes that we couldn't experience a de-vatting.[41]

Knowing what kind of kind a term picks out involves knowing what sorts of properties are *essential* to its application, and this is not something that can be conclusively determined *a priori*. For instance, I'm reasonably confident that 'vat' picks out *some* sort of interactional kind.[42] However, I could, in principle, be mistaken about this. One could imagine it turning out that, upon more careful investigation, all of the 'vats' we had ever encountered were living beings that moved around when they thought themselves unobserved, and that produced offspring that looked nothing like 'mature' vats. (These 'baby vats' were hidden away in 'vat factories' which were really just secret installations for the vats to grow up in.)[43] If this turned out to be the case, we might conclude that what 'vat' picked out was, after all, a *natural* rather than an interactional kind.[44] The 'baby vats' would still be vats even if they were small and

---

41 Which is more than just assuming that we *won't* experience a de-vatting. The relevant dispositions are present as long as de-vatting is possible.

42 Precisely what sort of kind it may be is less clear. It may, for instance, be an *artifact kind* term, whose deliberate construction to play a certain role in our lives is essential to its being a member of the kind. In such a case, there could be no 'naturally occurring' vats. Treating 'vat' as a term for an artifact kind rather than for the more generic sort of interactional kind suggested above would still, however, allow the term to be applied cross-environmentally.

43 This example is, of course, an adaptation of Putnam's own discussion of 'pencil' ('The Meaning of "Meaning,"' 242-3).

44 Or one could take the more extreme position that such a case would amount to our discovering that there were no, and never had been any, vats. For something like this view, see J. Katz, 'Logic and Language: An Examination of Recent Criticisms of Intentionalism,' K. Gunderson, ed., *Minnesota Studies in the Philosophy of Science VII* (Minneapolis: University of Minnesota Press 1975). However, Putnam himself clearly seems unsympathetic with this approach to such cases (see 'The Meaning of "Meaning"'). More plausibly, one might think that the term should be understood

couldn't hold liquids, and any vat-like object that *we* went on to construct would just be a mass of metal that looked like a vat.

The kinds of kind that our terms ultimately pick out will depend to a large extent on what kinds of kind 'work' best with our past, current, and future experience. We may, for instance, have originally taken 'air' to be a natural kind term of a sort that we still take 'water' to be, but such an understanding of the term proved unworkable. Experience has a way of 'boiling over' our current understanding of our environment, and an understanding of our terms that seems adequate at a time may have to be radically changed as our experience unfolds.[45] Many might find it plausible to treat 'vat' as a kind that applied exclusively to objects that could play a given role within our *currently experienced* environment, but such an understanding might collapse quickly if we suddenly found ourselves 'de-vatted.' Our current experience simply may not settle just what kind of kind 'vat' is.

The fact that we have only dealt with our current experiential environment can lead us to assume that having a 'physical' make-up is essential to being a vat. Nevertheless, having a 'physical' instantiation may be no more essential to being a vat than being white is essential to being a swan. While we typically use 'vat' to refer to vats in *this* experiential environment, future experience (involving either descending into virtual worlds, or emerging into a 'realer' world) might make our current assumptions about vats' 'physical' instantiation seem unessential to the term's meaning.[46] I know of no virtual worlds, and 'All vats are physical

---

as indeterminate between the natural and the interactional kind in the way that a term like 'dog' might be understood as indeterminate between an 'evolutionary' and a 'genetic' understanding. The latter of these would allow a 'synthetic dog' which had no genealogical connection to our dogs, but an identical physical and genetic make-up, to be a dog, while the former would not. See H. Putnam, 'Aristotle After Wittgenstein,' *Words and Life* (Cambridge: Harvard University Press 1994), 76-7.

45 The echo from W. James, *Pragmatism* (1907; Cambridge: Harvard University Press 1975), 106, is found in Putnam's work as well — see H. Putnam, *Pragmatism, An Open Question* (Cambridge: Blackwell 1995), 8 — and its relation to some of the views presented here are developed further in H. Jackman, 'James' Pragmatic Account of Intentionality and Truth,' *Transactions of the C.S. Peirce Society* **34** (1998).

46 One can see this in recent definitions of 'life,' where the properties essential to the kind are all of a functional/interactional sort that can be shared by various 'objects' found within the running of an appropriately programmed computer. See, for instance, the discussion of 'artificial life' in S. Turkle, *Life on the Screen* (New York: Simon and Schuster 1995). Furthermore, one might argue that the experience of 'devatting' would lead us to reshape our conception of the 'physical' in a way that

objects' may be true. Nevertheless, it may be true partially because of the way the world turns out to be, not simply in virtue of the fact that we take certain properties to be essential to being a 'vat.' What kinds of kinds our terms ultimately pick out can be neither infallibly determined by introspection, nor conclusively settled by convention. Indeed, this is something that Putnam himself has stressed more than just about anyone else.[47] By requiring that we treat certain current beliefs or presuppositions as essential to a term's meaning (and thus as unrevisable in the face of future experience without changing the meanings of the terms involved), Putnam's own argument thus presupposes a type of 'semantic essentialism' at odds with his generally 'pragmatic' picture of thought and utterance content.

Putnam's argument thus requires that we have *a priori* knowledge of the kinds of kinds that our terms pick out, and there is little reason to think that we must have such knowledge. The potential failure of semantic self-knowledge involved here is more robust than that relating to criticisms of (iii) and (iv). While we typically are not mistaken as to whether we are thinking or not, we often *are* mistaken about what properties are *essential* to the application of our terms.[48] Consequently there is nothing unintuitive in suggesting that we occasionally lack this sort of self-knowledge.

VII Skepticism and Metaphysical Realism

Nevertheless, even if the brain in a vat hypothesis is a coherent one, the suggestion that we are radically mistaken about the world (in the sense of having mainly false beliefs) still seems hard to defend from within an externalist framework. The 'transcendental' and 'magical' conceptions of reference that Putnam criticizes would allow a brain in a vat to have a term 'vat' which referred to vats but *didn't* refer to the 'vats' it experienced. By contrast, the position outlined here suggests that, while the reference of one's terms can extend *beyond* the sources of one's experi-

---

would allow 'natural kind' terms, and even the term 'physical' itself, to apply cross-environmentally.

47 See, for instance, his discussion of 'cat,' 'energy,' and 'pencil' in H. Putnam, 'It Ain't Necessarily So,' *Mathematics, Matter and Method* (New York: Cambridge University Press 1975), 'The Analytic and the Synthetic,' *Mind, Language and Reality*, and 'The Meaning of "Meaning."'

48 Indeed, our fallibility with respect so such questions has been evident in philosophic discourse from Socrates down to the present day.

ence, it typically cannot be *divorced* from them. The brain in the vat's term 'vat' is, after all, here taken to be instantiated in *both* the experienced and the unexperienced domain. A reinterpretation of one's current experience in the light of future experience typically will still leave the reanalyzed kinds applying to most of the currently experienced objects. The possible truth of the brain-in-a-vat hypothesis thus cannot be used to establish any sort of global skepticism of the sort that suggests that all of one's beliefs might be false. One may, for instance, still know that, say, one is looking at an apple, without knowing that one is not a brain in a vat because, even if one were a brain in a vat, one's claim to be looking at an 'apple' would still be true. The brain in a vat hypothesis is not a 'relevant alternative' that must be ruled out to be assured that one's claims are true.[49] Consequently, the position defended here is still 'anti-skeptical' to the extent that it suggests that even if we were brains in a vat, most of our beliefs about the world we experience could still be true.[50]

This would not, however, be enough to satisfy Putnam. The mere assurance that (even if we were brains in a vat) most of our beliefs would be true still leaves room for a considerable amount of epistemic disquiet. Brains in a vat, even if they typically have true beliefs, are fundamentally out of touch with reality's ultimate structure. In this sense, they still are 'radically mistaken' about the world. It is this worry that may ultimately be the target of Putnam's argument. After all, while Putnam admits that the possibility of our being brains in vats is typically used 'to raise the classical problem of skepticism with respect to the external world in a

---

49 Once again, this is assuming that 'brain in a vat' stands for the second scenario Putnam describes, which does not involve recent envatting or 'switches' between the 'real' and 'virtual' environments.

50 Of course, while the view makes room for the assurance that most of our assertions are true, by allowing that we cannot tell whether or not we are a brain in a vat, it may leave us open to a type of skepticism about our knowledge of the content of such true assertions. If it turns out that we don't, on such a view, know what we are saying, then it would be hard to claim that any of these true assertions amount to *knowledge*. There thus seems to be room for some sorts of skepticism here, even if it is not of the traditional 'all of my beliefs might be false' variety. (For a useful discussion of these issues, see Hymers.) Fortunately, I think that such worries about our knowledge of the content of our assertions and thoughts can be addressed. I have no space to do so here, but see H. Jackman, 'Semantic Norms,' 'Deference and Self-Knowledge,' *Southwest Philosophy Review* **16** (2000), and 'Ordinary Language, Conventionalism, and *A Priori* Knowledge,' *Dialectica* (forthcoming); for a related discussion, see D. Davidson, 'A Coherence Theory of Truth and Knowledge,' LePore, ed., and 'Knowing One's Own Mind.'

modern way,'[51] he claims that the possibility would be of interest only as a sort of 'logical paradox' if it were not for the sharp way in which it brings out the difference between 'metaphysical' and 'internal' realism.[52] The brain in the vat is supposed to illustrate the metaphysical realist's worry that even our best theory could be radically out of touch with the world's fundamental structure, and Putnam's argument is meant to show how this worry, characteristic of metaphysical realism, is incoherent.

Many have found this attack on metaphysical realism unconvincing, and the following passage from Wright is a typical (if unusually clear) expression of the intuition that Putnam's argument shows, at best, that we may not even be able to state how bad our epistemic position is:

> 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*.... But the sort of dislocation whose possibility is arguably implicit in metaphysical realism does not involve that its victims can conceptualize 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.... the real specter 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 specter is that of a thought whose truth would make a mockery of humankind 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.[53]

Putnam, in his reply to Wright, argues that this more abstract worry is, while perhaps less obviously so, as incoherent as the original assumption that we are brains in a vat.

> It is, perhaps, the vagueness of terms like "fundamental categories," "real kinds," etc., that conceals from Wright the fact that he is tacitly assuming conceptual access to such general notions as "physical" and "causation." But I take it that what we mean by "fundamental categories" and "real kinds" is kinds and categories that play a fundamental role in the description of physical things and their causal

---

51 *Reason, Truth and History*, 6

52 *Reason, Truth and History*, 49. For a discussion of this, see, once again, Davies, 'Putnam's Brain-Teaser' and Hymers.

53 Wright, 'On Putnam's Proof,' 239-40. See also Forbes, 'Realism and Skepticism.' The same sort of intuition, though more explicitly tied to the traditional problem of skepticism, is expressed in Nagel, 73.

> relations; if not, then I will ask Wright to give me an account compatible with externalism of how a being whose position is analogous to that of a brain in a vat would refer to the property of being "fundamental."[54]

Putnam's reply may indeed work as a response to Wright. Wright accepts Putnam's contention that a brain in a vat could not refer to vats, so he is not well placed to claim that it could refer to causation and fundamental categories. Nevertheless, Putnam's response to Wright ultimately runs into the same sort of trouble as his original argument. As with the case of 'cause,' even if the brains in the vat did not mean quite what *we* did by 'fundamental' (which I doubt), it is quite plausible to think that their term could be interpreted so that it applied to the world outside the vat as well.

The 'fundamental categories' for the brains in the vat should be the categories by which they could ultimately explain their experiences. If they were de-vatted, they would thus come to view (justifiably) certain categories relating to the computer as being 'more fundamental' than the categories they used before. Truths about the computer would, after all, explain why certain apparently 'fundamental laws' in their virtual environment were true. 'Fundamental categories' apply to the total range of potential experience, and are not limited to the experiences available at a particular point. A brain in a vat can refer to categories in our world by its expression 'most fundamental' because *future* experience could (though not necessarily 'will') reveal them to be so.[55] This is why the brains could describe the world outside the vat as 'more real' than the one they are currently experiencing. Putnam wishes to understand truth and reference in terms of rational acceptability under 'epistemically ideal conditions,'[56] and there is no reason why (given that a de-vatting is physically possible)

---

54 'Comments and Replies,' 287-8

55 Putnam seems to suggest that a future de-vatting could only be relevant by allowing the brain in the vat access to descriptions such as 'the things I will refer to as "vats" at such and such a future time' (*Reason, Truth and History*, 16). In much the same way, Putnam claims that the brain in a vat hypothesis would be a coherent one if it predicted a de-vatting some time in the future (*Reason, Truth and History*, 131). However, even the *potential* for de-vatting, whether it is actualized or not, is relevant to the interpretation of one's terms. Even if the brain is not de-vatted, it is still *disposed* to respond to its de-vatting in a particular way, and these dispositions help constitute what it should be interpreted as meaning by its terms.

56 *Reason, Truth and History*, 55. Or at least the Putnam of *Reason, Truth and History* did. Putnam's views on the topic of truth have changed since then. See H. Putnam, *Representation and Reality* (Cambridge: The MIT Press 1988) and 'Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind,' *Journal of Philosophy* **91** (1994).

potential out-of-vat experiences should be excluded from these 'ideal conditions.'[57]

Putnam's intuitions about the incoherence of the brain in a vat scenario may be driven by a misplaced assimilation of it to the metaphysical realist's worry about our never getting hold of things as they are in themselves. In formulating the metaphysical realism/internal realism contrast in the way that he does, Putnam seems to be trying to use the brains in a vat to give a 'naturalized' version of a Kantian phenomena/noumena distinction. The vat's virtual world would here takes the place of the world of experience, and the world outside of the vat would play the role of the world as it is in itself. This understanding of the brain in a vat's predicament is suggested by Putnam's initial formulation of the problem in his 'Realism and Reason.' In that paper, he asks how, if we were brains in a vat, would it come about 'that *our* word "vat" refers to *noumenal* vats and not to vats in the image?'[58] However, unlike Kant's noumena, experience of the world outside the vat is not, *in principle*, inaccessible to the brains in the vat.[59] The world outside the envatted brains is beyond the reach of their current experience, but there is nothing necessarily *inexperientiable* about it. The world outside the vat is not 'the world as it is in itself.' It is a world that a brain in a vat could (but unfortunately doesn't) experience. If it were de-vatted (and given its new and expanded range of experience), the former brain in a vat would rightly deny that the theory of the world developed in the vat

---

57 As Horgan puts it, the relevant sense of idealization must also include an idealization of the cognizer's epistemic vantage point, T. Horgan, 'Metaphysical Realism and Psychologistic Semantics,' *Erkenntnis* **34** (1991), 303. That such a qualification is implicit in Putnam's own view of truth is suggested in D. Davies, 'Why One Shouldn't Make an Example of a Brain in a Vat,' *Analysis* **57** (1997), and Putnam makes it more explicit in H. Putnam, *Representation and Reality*, 'Reply to Terry Horgan,' *Erkenntnis* **34** (1991), and 'Sense, Nonsense, and the Senses.' For a discussion of this aspect of Putnam's recent writings on truth, see C. Wright, 'Truth as Sort of Epistemic: Putnam's Peregrinations,' *Journal of Philosophy* **97** (2000).

58 'Realism and Reason,' 127 (Italics, as elsewhere, are Putnam's). Note that not only are the unexperienced vats outside the vat referred to as 'noumenal,' but it is also assumed that the hypothesis requires that vats in the image do *not* fall within the extension of 'vat.'

59 Exactly how Kant's distinction between appearances and things in themselves should be understood is a notoriously difficult topic. Nevertheless, it seems fairly certain that, whatever the proper understanding of Kant's distinction between the noumena and phenomena is, it will be significantly different from the relation between experiences within the vat and experience of the world outside of it. For instance, the noumena are not distinct objects from the phenomena that could, in fact, eventually be objects of experience themselves.

meets its 'highest cognitive standards,' since accounting for past, current, and future experience is manifestly such a cognitive standard. A brain in a vat might rationally inquire indefinitely without discovering its predicament, but 'the best theory possible' for the brain in a vat would not be one that left out the world external to the vat, since experience of that world is at least potentially available to it.[60] Saying that we might be brains in vats is, after all, compatible with saying that we could, in principle, come to *recognize* that we had always been brains in vats. Even *if* the fundamental features of the world must be *experienceable*, they need not be *experienced*.[61]

VIII Conclusion

Putnam considered the brain in a vat hypothesis to be philosophically significant because he took the purported self-refuting character of this sort of skeptical worry to undermine the plausibility of metaphysical realism. However, Putnam's argument against the possibility of our being brains in a vat relies upon treating the terms in the 'skeptical' hypothesis as if they picked out sortals that could not be applied cross-environmentally. There is, however, no compelling reason to think that such an assumption, even if true, could be established *a priori*. Consequently, there is no way to establish that a brain in a vat couldn't truly think that it was brain in a vat. It may thus be, in some sense, possible that we are all brains in a vat. Nevertheless, the possibility that we are brains in a vat, so understood, supports neither the skeptic's suggestion that most of our beliefs could be false, nor the metaphysical realist's worry that the best theory we could *possibly* come up with might still be radically out of touch with the world's fundamental structure. The coherence of the brain in a vat scenario simply does not have the philosophical implications that Putnam fears, and semantic externalism may thus lack many of the metaphysical and epistemic consequences that Putnam hoped for it. Since such purported metaphysical and epistemic consequences provided many with good reason to be wary of

---

60 Say, the 'de-vatting' also occurs through some 'cosmic' coincidence. If the first coincidence is 'physically possible' then it should seem as if the second would be as well. There should, then, be no problem tying such potential experiences to idealized, if not actual, inquiry.

61 No commitment need be taken on here about the truth-value of this conditional's antecedent.

semantic externalism, such a result should ultimately make the view more plausible.

Received: July, 2000
Revised: February, 2001