"Eligibility and Ideology in the Vat"[1]

Tim Sundell
University of Kentucky

forthcoming in Sanford Goldberg (ed.)
Putnam on Brains in a Vat
Cambridge University Press

draft of October 5, 2014

Introduction

Lewis's *reference magnetism* is meant to address worries raised by Putnam's anti-realist *model theoretic argument* (MTA). I won't try to determine whether it succeeds in that regard. But suppose—as many contemporary metaphysicians do—that it does succeed. Putnam's got another argument. Putnam's *brain in a vat argument* (BVA) is not just another attempt to respond to Cartesian skepticism. It is *also*, like the MTA, an attack on realism. And because the brain in a vat argument proceeds on the basis of different—and more conservative—assumptions about the nature of reference, a solution to the MTA does not automatically extend to the BVA. So: if we suppose that reference magnetism succeeds in addressing the worries raised by the MTA, what should we think of its prospects for mounting a response to the BVA? I argue that the metaphysical realist does have a response to the BVA. That response, however, is importantly different from what the realist might have thought going in. I argue that the realist should insist on a distinction between a theory's *truth* and its *overall epistemic success*. In turn, the realist can maintain that there *are* genuinely radical yet non-self-refuting skeptical hypotheses, but that such hypotheses concern not the truth of a theory but a different aspect of epistemic success, namely the fundamentality of its ideology. Such a response is consistent with the conclusion of the BVA. Nevertheless, it deprives that argument of its anti-realist force. The view I suggest thus makes theoretical space for the semantic considerations Putnam brings to bear while preserving the spirit of the metaphysical picture he attacks.

[1] Sincere thanks for helpful discussion and comments to Mahrad Almotahari, Clare Batty, Gunnar Björnsson, Josh Brown, Andy Egan, Tim Fitzjohn, Sandy Goldberg, Jennifer Hudgens, James Lincoln, Rebecca Mason, Adam Patterson, David Plunkett, and Meg Wallace. I owe a special debt to Tim Button, without whose work and patient correspondence this essay could not have been written. Mistakes, confusions, and outrageous over-simplifications are of course my own.

Section 1: Magnets and Model Theory

Reference magnetism is the view that part of what determines the meanings of our terms is the objective metaphysical naturalness of the properties and relations to which we refer. Lewis introduces the view (though not under that name) in the context of arguing that there is a great deal of philosophical work to be done by such a notion of naturalness—by the notion that the world contains metaphysical joints the location and nature of which are independent of us and our theorizing.[2] The world can be carved up in all kinds of ways. But Lewis argues that some ways of carving the world are objectively, metaphysically, more natural than others. He argues for this claim by demonstrating some of the philosophical work it can do, and an important bit of that work concerns the determination of the contents of our language and thought. While there are many possible interpretations of my language—many possible mappings from my predicates to various properties and relations—only some of those interpretations will ascribe reference to properties and relations that are metaphysically natural, as opposed to properties and relations that are gerrymandered and perverse. Of course we could refer to gerrymandered categories if we tried. And a full development of the view will acknowledge that many of our everyday concepts pick out categories that are far from metaphysically fundamental.[3] But those horribly, hopelessly gerrymandered categories—the grues, the quuses, the coins-in-my-pocket-plus-Eiffel-towers, the cow-or-electrons—those simply aren't eligible to be the kinds of things to which we refer with our ordinary terms. Not when alternative interpretive schemes are available that are simpler and more charitable.

Lewis introduces this view in an attempt to defend the metaphysical realist from an anti-realist attack by Putnam. In particular, he's responding to Putnam's model theoretic argument (MTA). That argument, very roughly, goes like this. Consider an ideal theory T.[4] By "ideal," I'll just mean that T is fully predictively adequate and that it has all of the members of your favorite list of theoretical virtues. As Putnam puts it, T is "operationally ideal". If T is ideal, then it's consistent. If T is consistent, then there is a model—an interpretation—on which it is true.[5] Pick one such model and call it M. Now take the domain of M and map the objects in that domain one-to-one to objects in the world. Now reinterpret the predicates of T, grouping objects in the world not by how those objects seem to us to go together, but

[2] Lewis (1983).
[3] See Sider (2013), pp. 37-39.
[4] It's not crucial to this part of the argument that T be ideal—it just needs to be consistent—but it matters later.
[5] See Putnam (1980), p. 473, and Putnam (1977), p. 485. Williams (2007) discusses the difference between "model existence" arguments like the one I gesture at here and "permutation arguments," in which the correspondence between the terms of a theory and the entities in some domain is scrambled to preserve truth. As that work shows, there are important differences between the two types of argument, but since I set aside the MTA below to address the BVA, the differences will not matter to me here.

instead mechanically, in ways mirroring their correspondents' grouping on M. This procedure yields a second interpretation, N, on which T is not just true, but true of the world. Without ever having looked at what the world is *actually like*, we've constructed an interpretation on which theory T is not just operationally ideal, but true. In other words, for any (consistent) theory and any (big enough) world,[6] there is some word-world relation (indeed, many) on which that theory is true of that world.

The argument goes on. Suppose you object that N is not the intended interpretation of T. Yes, it makes T true of the world, but since it's not the *right* interpretation of T, that's neither here nor there. How is such a response to be defended? How are we to demonstrate that N is not the intended interpretation? Proposing further constraints on reference won't help. Anything we suggest or stipulate will itself be part of our own theorizing and thus open to the same referential rejiggering that gave us N in the first place. The rejiggered interpretation of our new, augmented theory will satisfy the proposed constraint on reference—whatever it may be—along with everything else. It also won't help to say instead that there are intrinsic relations—independent of our theorizing—between our words and their referents. That kind of view is tantamount to belief in the supernatural, to the antiquated thought that symbols are magically connected to the things they represent or that the world gives things their names. For Putnam, "we interpret our languages or nothing does."[7] But we can't rule out interpretation N. So nothing can.

The MTA is an argument for semantic indeterminacy, an argument about language and thought and reference. But why should we think of it as an *anti-realist* argument? The kind of realism Putnam attacks in "Models and Reality" and *Reason, Truth, and History*—what he calls *metaphysical realism*—is a view on which truth is "radically non-epistemic". (For convenience, I'll simply use *realist* to pick out the type of metaphysical realist who is the target of the MTA.) For the realist, what the world is like is a matter independent of us and our concepts and our theorizing. There is an objective way that the world is, complete with objects and properties and relations, and the question is simply how close our theories come to describing it accurately. For Putnam, the most important consequence of such a view is that even an operationally ideal theory could turn out to be hopelessly, radically mistaken. (After all, "we might all be 'brains in a vat,' the metaphysical realist tells us."[8]) Since, for the realist, there is no constitutive relation between ideal theory and truth—which is just to say that truth is non-epistemic—the two must be

[6] The argument can go wrong if the world is too small. I set this wrinkle aside here.
[7] Putnam (1980), p. 482.
[8] Putnam (1977), p. 485.

able in principle to come apart. Call this—the possibility of radical error in ideal theory—*pervasive error*.[9]

The dialectical position for the realist on this understanding is admittedly odd. It's hardly the case that all metaphysical realists are skeptics after all. But the point is not that the realist must accept skepticism itself. The claim is not that the realist must accept, when all is said and done, that she could not know her own best theory to be true. The point is just that she can *make sense* of the possibility that it is not, that "the theory that is 'ideal' from the point of view of operational utility, inner beauty and elegance, 'plausibility', simplicity, 'conservatism', etc., *might be false*".[10] Whatever the realist's favorite anti-skeptical maneuver, the skeptical scenario itself is a coherent possibility, one to be taken seriously and addressed with care. Since the skeptical scenario is coherent, so is pervasive error. And since pervasive error is coherent, the realist can maintain her radically non-epistemic view of truth.

The problem is that if the MTA is successful, then pervasive error is not a coherent possibility. For an ideal theory to be false, it must be the case that the theory fails to correspond to what the world is like *on the correct interpretation of that theory*. But the MTA shows that there is no way to privilege such an interpretation as correct. The theory is guaranteed to be true on some interpretation, and nothing from inside *or* outside of the theory can show that that interpretation is the wrong one. (Everything *inside* of it is "just more theory," and nothing *outside* of it is the kind of thing that can interpret our theories. Only we get to do that.) The best thing to say is that it was a misguided project to think of truth as independent of ideal theory in the first place. That's what required us to drive a wedge between the two. If we simply define truth in terms of ideal theory—if we adopt an epistemic notion of truth—then the problem doesn't arise. But of course, if we define truth in terms of ideal theory, then the thought that such a theory could be false is incoherent. Thus, and not to put too fine a point on it: If metaphysical realism is true, then pervasive error is a coherent possibility. But pervasive error is not a coherent possibility. So metaphysical realism is false.

Lewis's response gets around Putnam's argument—if it does—by avoiding what Lewis calls the "just more theory trap".[11] Lewis rejects Putnam's declaration that we interpret our language or nothing does. We have a role to play, of course, but there are other factors. If we look to the referents, in addition to the referrers, then we will see that "reference consists in part of what we do in language or thought when we refer, but in part it consists in eligibility of the referent. And this eligibility to be referred to is a

[9] Button (2013) gives to this possibility—or rather, to worry about the possibility's obtaining—the excellent label *Cartesian angst*. Lewis (1984), p. 235 refers to Putnam's target simply as *massive error*. For reasons that will become clear below, I prefer a label closer to Lewis's, but I think there is not a deep philosophical difference here.
[10] Putnam (1977), p. 485. Emphasis Putnam's.
[11] Lewis (1984), p. 228.

matter of natural properties."[12] The correct interpretation of a language is the one that assigns eligible referents—properties and relations that are reasonably metaphysically natural—to the predicates of that language. This is not the claim that "eligibility-theory is to be satisfied somehow," or that "the referents of 'cat' etc. are to be included among the referents of 'eligible'."[13] That type of view is clearly vulnerable to the "just more theory" objection. But Lewis's suggestion is meant to be different. It's not our theorizing about the joints in nature that determines which referents are most eligible. It's the joints themselves. To accuse Lewis of simply doing more theory is to misunderstand the view.

Of course, if Lewis does evade the "just more theory trap," then he may still remain open to the accusation of magical thinking. Lewis quotes Putnam calling the view "spooky" and "medieval sounding" and asks

> [W]hat's wrong with sounding medieval? If the medievals recognized objective joints in the world—as I take it they did, realists and nominalists alike—more power to them. But I don't suppose that inegalitarianism of classifications is an especially medieval notion—rather, egalitarianism is a peculiarity of our own century.[14]

More recent defenders have gone further, arguing that in fact the substance of the eligibility view follows quite naturally from nothing more exotic than Lewis's commitment to simplicity as a theoretical virtue.[15]

If the view works, and if it thus makes sense to talk of a privileged interpretation after all, then it makes sense to ask, even of an ideal theory, whether it gets the world right or hopelessly, radically wrong.

> When we limit ourselves to the eligible interpretations, the ones that respect the objective joints in nature, there is no longer any guarantee that (almost) any world can satisfy (almost) any theory. It becomes once again a worthy goal to discover a theory that will come true on an eligible interpretation, and it becomes a daring and risky hope that we are well on the way toward accomplishing this.[16]

This hope—and the corresponding possibility of the hope's being dashed—is what saves realism from Putnam's attack. Interpretations that, like N, are constructed mechanically to guarantee truth will

---

[12] See Lewis (1983), p. 371.
[13] Lewis (1984), p. 228.
[14] Lewis (1984), p. 229.
[15] Sider (2013), Ch. 3, draws on Williams (2007) to make this argument. Williams argues that for Lewis, simplicity of a theory is measured by the number and syntactic complexity of the axioms required to state it. But to avoid triviality for such comparisons, what must be measured is the number and syntactic complexity of the the axioms when stated in a language the primitives of which pick out fundamental properties. A gruesome theory might at first appear equally syntactically simple as a non-gruesome theory. But when both are stated in primitive terms corresponding to metaphysically fundamental properties, the gruesome theory will require more, and more complex, axioms. The eligibility constraint for theories of interpretation in particular follows from this. Williams goes on to make trouble for the eligibility constraint, however, arguing that the constraint successfully addresses permutation arguments, but not model-existence arguments for indeterminacy. As noted above, that issue is outside the scope of this paper, in which I set aside the MTA in order to address the BVA.
[16] Lewis (1984), 227.

inevitably make reference to properties and relations that are highly non-natural, and thus ineligible to be the referents of our words. The truth of N really is neither here nor there. We want to know whether the theory is true on an *eligible* interpretation.

My topic here is not whether reference magnetism succeeds as a response to the MTA.[17] Whether Lewis's view avoids the "just more theory trap," and whether, if it does, it also avoids the accusation of magical thinking is not a settled matter. But many contemporary metaphysicians work in a realist vein and do not take the MTA to be a decisive consideration against doing so. Indeed, arguably many of those metaphysicians proceed in this way in part on the basis of their endorsement of reference magnetism. If the view is plausible enough for them, then it's plausible enough to take as a supposition in addressing related arguments downstream. The question of whether the realist has the resources to respond to Putnam's *other* anti-realist argument therefore has independent philosophical interest, even if we ask it under the potentially controversial supposition that the MTA has been successfully addressed.

Section 2: Reference in the Vat

A great deal of work has been devoted to reconstructing the best version of Putnam's BVA, but a recent and especially pointed reconstruction, due to Tim Button, will serve my purposes. Putnam (1981) begins by asking that we consider a very particular kind of skeptical scenario. We are brains in vats. We have always been brains in vats. Our brains are hooked up to computers and also to each other. (So when we take ourselves to have a conversation, we are in fact communicating.) But there is nothing in the universe other than the brains and the vats and the computers. In particular, there is no programmer. Let's say that by quantum fluke the universe came to consist of the brains and the vats and the computers and nothing else. This is what I'll call the BIV scenario.

Putnam famously asks as to reflect on what a brain in such a scenario refers to with its words. Consider, for example, the brain's word "cat". Very weak semantically externalist assumptions suggest that the brain's word "cat" doesn't refer to cats. After all, the brain has never encountered a cat in its life. It's never met anyone else who has either. It hasn't encountered the kinds of things a cat is made out of, and cats play no role in the causal processes surrounding the brain's use of its word. What, then, does the brain refer to with its word "cat"? Perhaps the brain refers to nothing; perhaps its predicament is so dire as to prevent it altogether from making claims or having thoughts with content. We don't have to think that though. There is something that the brain tracks with uses of its word "cat". It's just that the something isn't cats. What is it? Opinions vary, but I'll simply call it cats-in-the-software.

[17] See Williams (2007), Sider (2013) Ch. 3, and Button (2013) Ch. 3-4 for some recent discussion.

The brain has encountered cats-in-the-software many times, and its encounters with cats-in-the-software are systematically linked to the brain's tokening of its word "cat". When the brain says something using its word "cat," it's thus reasonable to think that what it's talking about is actually cats-in-the-software. When a brain B says "there's a cat in front of me," what it expresses, on this view, is a proposition like this: There's a cat-in-the-software standing in the in-front-of-in-the-software relation to B. And that proposition might very well be true. Over the course of the brain's life there has been plenty of cats-in-the-software that have stood in the in-front-of-in-the-software relation to it. Supposing this is one of those occasions, and given the contents of what it has actually said, the brain isn't wrong after all. Indeed, as we apply the argument more generally, we find that the brain is deceived about much, much less than we originally thought it was.

So far we've seen that the brain in a vat might not be as bad off, epistemically, as we thought. But that isn't the BVA. Or at least it isn't yet. The conclusion of the BVA is "I am not a brain in a vat" and that conclusion has not yet been reached. So how do we get to the strong conclusion from the considerations just described? Button (2013) offers the following reconstruction.[18] What goes for cats goes for brains,[19] so:

(1) A BIV's word "brain" does not refer to brains.
(2) My word "brain" refers to brains.
(3) I am not a BIV.

Not everyone will agree that the brain's word "brain" succeeds in referring to brains-in-the-software. But to accept premise (1), you just need to think that it fails to refer to brains. The only thing required to sign on for that is a commitment to some reasonable version of a causal constraint on reference. Putnam doesn't think such a constraint can reach outside of our theories to refute the MTA. But he certainly thinks such a constraint is in fact an internal part of our best theory of reference. Most people do. If you think reference has to do with and is in some way constrained by your environment—if, say, you think that residents of Twin Earth don't refer to H_20 with their use of the word "water," or if you think that Putnam's ant has failed to create an actual representation of Churchill with its random tracings[20]—then you should probably be willing to grant premise (1).

[18] See Button (2013) ch. 12, and this volume, ch X. In giving this version of the argument, Button gives credit to Tymockzko (1989), Brueckner (1992), Ebbs (1992), Wright (1992), Putnam (1992), Warfield (1998), and DeRose (2000).
[19] There are no cats in the brain's universe, while there are brains. But pretty uncontroversially, the brain's causal contact with itself is not the right kind of causal contact to underwrite reference.
[20] Putnam (1981), p. 1.

The only other thing the argument needs is disquotation.[21] Disquotation is not always entirely innocent as an argumentative move of course.[22] But the skeptic is hardly in a position to raise a fuss about its application in premise (2). After all, "even to understand or talk about about the BIV scenario at all, we need to rely on disquotation. Otherwise, the BIV scenario does not confront us with the worry *that* we are brains in vats."[23] In other words, the skeptic herself requires that our word "brain" refers to brains. If she challenges premise (2), then she calls into question whether she has succeeded in presenting us with the skeptical scenario.

The BVA, especially when formulated in this way, is a pretty good argument. Still, it is no more obvious than it was with the MTA why, *prima facie*, we should take it as an *anti-realist* argument. Shouldn't the realist be happy? Putnam insists that the realist must treat pervasive error as a genuine problem. That means she needs a solution—and now she has one. In fact, though, the anti-realist application of the BVA is the same as the anti-realist application of the MTA. Both arguments attack the coherence of pervasive error. The MTA concludes that pervasive error is incoherent by showing that nothing could force us to interpret an ideal theory in such a way as to make it false. The BVA concludes that pervasive error is incoherent by going directly after the realist's method for demonstrating its coherence.

The realist demonstrates the coherence of pervasive error by observing that we might be brains in vats. The BVA—in contrast to the MTA—raises no questions about whether there could be a determinate interpretation of the brain's language. But it asks us to observe that very conservative constraints on reference show that such an interpretation simply *isn't* one on which the brain's theory is massively mistaken. That's the point about how a brain isn't as bad off, epistemically, as we might have thought. More importantly, the BVA demonstrates the the brain in a vat *lacks the representational resources to describe its own predicament*. The brain in a vat, in other words, cannot contemplate the BIV scenario. That means that if I can contemplate the BIV scenario, then I'm not in it. The possibility that the realist mobilizes to demonstrate the coherence of pervasive error is such that simply to contemplate it is to know that it doesn't obtain.[24] Once again, if metaphysical realism is true, then pervasive error is a coherent possibility. But pervasive error is not a coherent possibility. So metaphysical realism is false.

---

[21] Putnam (1992), p. 369: "The premises of the Brain in a Vat argument are (1) the disquotation scheme for reference […] and (2) that reference to common objects like vats, and their physical properties […] is only possible if one has information carrying causal interactions with those objects and properties, or objects and properties in terms of which they can be described."

[22] Wright (1992) argues that even if semantic externalism denies us certain kinds of semantic self knowledge, the knowledge it denies us is not what's needed for disquotation of the kind in premise (2).

[23] Button (2013), p. 125. Button attributes this argument to Tymoczko (1989).

[24] It is notoriously difficult to state with precision what is wrong with the BIV scenario according to the BVA. It is in some sense "self-undermining," a sense that I won't attempt to make more precise. For me it will be enough to

Section 3: Eligibility in the Vat

What exactly should the realist say in response to the BVA? Lewis himself does address the vat argument in his (1984). But he treats the argument as an attempt at "exonerating" the brain of error.[25] In other words, he takes the conclusion to be that the brain is less bad off, epistemically, than we might have thought. Lewis argues in response that Putnam overestimates just how far the argument generalizes, and that therefore the brain may have a great many false beliefs after all. Setting aside issues of Putnam interpretation, however, this kind of response won't work against the reconstruction I discuss here. Again, the conclusion of what I've called the BVA is not "being a brain in a vat wouldn't be so epistemically bad after all". Rather, the conclusion is "I'm not a brain in a vat." And this version of the BVA doesn't require that the argument generalize at all. That the brain's word "brain" doesn't refer to brains is enough to do the trick.[26]

So does the realist have available another form of response? It's important at this point to recall that there are two arguments at work here. There is the BVA itself, which concludes that I am not a brain in a vat. And then there is the anti-realist modus tollens, which concludes that metaphysical realism is false. The connection of course is that the modus tollens gets support for its second premise—"pervasive error is not a coherent possibility"—from the BVA. The realist could thus mount a defense either by addressing the BVA itself or by taking on the anti-realist modus tollens. I consider the latter option below, but first I consider how the realist might address the BVA itself. That is to say, how the realist could defend the claim that I might, after all, be a brain in a vat. The prospects for a realist response of this type will of course depend greatly on what, specifically, the realist has to say about the brain and about the reference of its terms.

The first thing to note in this regard is that—*prima facie* at least—the realist's acceptance of reference magnetism provides nothing in the way of special or surprising results. Reference magnetism

---

focus on the observation that the brain lacks the representational resources to describe its own predicament, and to ask whether the same holds for various other skeptical hypotheses.

[25] Lewis (1984), pp. 234-236.

[26] Chalmers (2005) offers a take on the brain in a vat that, though it differs in emphasis, bears similarities to Lewis's response. Chalmers suggests that the brain can be exonerated quite extensively of error, but that a number of its more fundamental metaphysical beliefs are false. The brain believes, for example, that it lives in a universe in which fundamental physical particles are not themselves constituted by a yet more fundamental layer of computational processes. But, since the things the brain calls "tables" are made out of bits of software, that belief is false. The skeptical scenario is thus, for Chalmers, a collection of metaphysical views rather than a distinctively nightmarish epistemic predicament. Chalmers's view, like Lewis's, combines agreement with Putnam that the brain's beliefs are true when it comes to ordinary objects with the contention that the argument will at some crucial point fail to generalize. (In that essay, Chalmers also employs an "exemplification" response to Putnam's argument. See Section 5 below for discussion of that strategy.)

was brought in to address radical indeterminacy of semantic content. It allows the realist to reject perverse interpretations of our language. But the BVA never raised the possibility of radical indeterminacy of semantic content. All it asked us to accept was some kind of causal constraint on reference. And it never raised the specter of perverse interpretations either. Indeed, the interpretation that is tentatively proposed in the course of presenting the argument—cats-in-the-software, in my version—is not just determinate but highly charitable. Lots of people who reject the set up for the MTA—"we interpret our language or nothing does," in particular—are going to accept a causal constraint on reference. That puts the realist in pretty much the same boat as everyone else.[27]

So what should the realist say about the brain's words?[28] If the realist concludes that the brain fails to refer, then the BVA will immediately go forward: If the brain's word "cat" doesn't refer to anything, then it doesn't refer to cats. And that, plus disquotation, is all that's required. So the realist will have to argue that the brain does succeed in referring. What specifically do the brain's words refer to? Well if the brain's word "cat" refers to cats-in-the-software, then the BVA *still* goes through. (Since referring to cats-in-the-software is one way of not referring to cats.) So: if the realist is to defend the possibility that we might be brains in vats, she will have to argue (1) that the brain's word "cat" succeeds in referring and (2) that in particular it succeeds in referring to cats.

At this point in the argument, it might be tempting to lean—heavily—on the eligibility constraint. The property of being a cat is more natural than the property of being a cat-in-the-software. To see this, just consider that as strange as the BIV scenario is, it takes place in a universe composed of the same fundamental parts, governed by the same fundamental laws, as our own. There *are* ordinary physical objects in the BIV scenario. (Brains, vats, computers, wires, etc.) It just so happens that the brain never interacts with them as we do. A complete scientific account of the BIV's universe will thus, at the fundamental level, look much like an account of our own. And a fundamental theory of our own universe is going to take a long time before it gets around to describing the contents of computer programs. It will have to tell us what a computer is first, for one thing, which means it will already have worked its way up to medium sized dry goods. Once it's done that, it will already be able to tell us about cats. Consider also

---

[27] Sider (2013), Ch. 3 points out that Lewis's eligibility constraint can be combined with a variety of other views about reference. You could use it to augment a kind of global descriptivism, as Lewis does for the sake of argument in presenting the view. But you could just as easily combine it with other metasemantic views. But, whether you hold to global descriptivism, or the causal historical theory, or Millikan's metasemantics, or conceptual role semantics, or whatever, you're still likely to think that residents of Twin Earth do not have H_20 in the extension of their word "water." And if you think that, then you're unlikely to object to the semantic assumptions at work in the BVA.

[28] It is no longer important for the argument that we talk about the brain's word "brain" in particular, and doing so makes the discussion sound unnecessarily baroque. I switch the example back to "cat" now, simply so that I don't have to use the word "brain" so much. Everything I say should apply to brains and to "brain" as well.

that, as Fodor famously observes, *content* itself is unlikely to appear in the "complete catalogue […] of the ultimate and irreducible properties of things".[29] So before the theory can tell us about the contents of software, cats-in-the-software included, it will first have to explain intensionality. And to do that, again, it will have to have worked its way up to ordinary objects and organisms—like brains and computers—since electrons and quarks are not the kind of things that are capable of referring. If the theory has reached a point where it is working in terms of categories like *brain* and *computer* and *aboutness*, then it will long since have reached a stage where it can make sense of a category like *cat*.

Reference magnetism is precisely the view that an interpretation mapping more natural properties to our terms is to be preferred over an interpretation mapping less natural properties to our terms. As we've just seen, the property of being a cat is more natural than the property of being a cat-in-the-software. So perhaps the realist could declare cats-in-the-software *ineligible* as a referent, and cats—even though they don't happen to exist in the brain's world—*eligible*. Thus, the brain's utterance of "there's a cat in front of me" would mean that there is a *cat* in front of it, which is false. Precisely what the realist wants to be able to say.

This maneuver is actually in keeping with other uses to which reference magnetism has been put. In his (2003), Weatherson puts the view like this:

> [F]or any predicate t and property F, we want F meet two requirements before we say it is the meaning of t. We want this meaning assignment to validate many of our pre-theoretic intuitions […] and we want F to be reasonably natural […]. In hard cases, these requirements pull in opposite directions; *the* meaning of t is the property which on balance does best.[30]

Weatherson's point is that reference magnetism allows us to maintain some distance between pre-theoretic intuitions and meaning. Here's one of his examples of how that could work: For eighteenth-century speakers who used their word "fish" to describe whales, an interpretation on which that word really meant *fish-plus-whales* would vindicate more of their intuitions. But an interpretation on which their word "fish" simply meant *fish* would still respect *most* of their intuitions, and it would *also* respect the facts about naturalness. It is sensitive to speaker usage and intuition, but still allows us to say that an

---

[29] Fodor (1997), p. 97. The complete quote is: "I suppose that sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, the likes of spin, charm, and charge will perhaps appear upon their list. But aboutness surely won't; intentionality simply doesn't go that deep."

[30] Weatherson (2003), p. 9. Emphasis Weatherson's. Those philosophers who combine the eligibility constraint with different metasemantic theories may object to Weatherson's suggestion for this other requirement. But analogous considerations apply. For example, if eligibility is combined with Millikan's metasemantics rather than Weatherson's pre-theoretic intuitions constraint, we will see that one interpretation but not the other assigns to the brain's word "cat" an extension that neither it nor any of its ancestors has ever successfully *tracked*, in Millikan's sense. See Millikan (1987). That makes the property of being a cat a terrible interpretation of the brain's word "cat," and it does so for reasons analogous to those described above in terms of Weatherson's suggestion.

eighteenth-century utterance of "whales are fish" expresses the false proposition that whales are fish, rather than the true proposition that whales are fish-or-whales. The first interpretation hews too closely to speaker usage, and doesn't allow us to make sense of the fact that the speakers, however consistent their usage of the term "fish," were in error about whales. The second interpretation thus represents the better balance of the two constraints.

In the case of the BIV, as in the case of Weatherson's eighteenth-century mariners, we are trying to make sense of systematic error. The brain's case, however, is simply too extreme for this kind of application of the eligibility view. As the passage from Weatherson makes clear, the eligibility constraint is meant to be one consideration among others in choosing an interpretation. Cat-hood is indeed a more natural property than cat-in-the-softwarehood, just as being a fish is a more natural property than being a fish-or-whale. But an interpretation on which a BIV's word "cat" refers to cats respects *none* of the brain's pre-theoretic intuitions. It correctly describes none of the patterns in the brain's usage of that word, and accurately characterizes none of the brain's linguistic dispositions.

Moreover, cat-in-the-software is not itself an entirely unnatural property. It's more natural, for example, than cat-or-Eiffel-tower-in-the-software or cwat-in-the-software. (We'll let cwats-in-the-software be a property possessed by cats-in-the-software along with lamps-in-the-software if it's a Thursday-in-the-software.) The categories needed to explain the brain's linguistic behavior and pre-theoretical intuitions will have a role to play in a theory of the brain's world. Cat-in-the-software is one such category. Cwat-in-the-software is not. So cat-in-the-software is a less natural category than cat, but (a) it is still reasonably natural and (b) an interpretation assigning it to the brain's word does a much better job of meeting other constraints on reference than one assigning cats. However great the difference in naturalness, eligibility is to be balanced against some degree of charity or sensitivity to usage.[31] Using the eligibility constraint to force a *cat*, rather than *cat-in-the-software*, interpretation privileges naturalness to the thorough exclusion of any of the other factors that play a role in the determination of reference. No plausible version of reference magnetism could get—or would want to get—that result.

The point here is that at the level of first-order semantics, the realist and Putnam will have largely the same kinds of things to say about the semantic values of various speakers' terms. As emphasized above, most advocates of reference magnetism are going to accept some kind of causal constraint on reference. And—in stark contrast to the MTA—some kind of causal constraint is really all that's needed for the BVA to have bite. As Weatherson argues, the eligibility constraint can be used to pull questions about meaning apart from questions about dispositions or first-order intuitions. It does help us explain how we could get things consistently wrong. But while that may be a plausible analysis for errors like

[31] Or attention to the properties actually being tracked, or to conceptual role, etc.

thinking mistakenly that whales are fish, it's far less plausible for more radical and wide-ranging forms of error. As long as the eligibility constraint is to be balanced against other factors—causal contact, dispositions, intuitions, tracking, etc—it can't be strong enough to wrestle the brain's reference all the way out of the vat and onto the uninstantiated property of cathood. Such a view would count as magical, even by the lights of the realist. So what should the realist say about the brain's word "cat"? Pretty much the same thing everybody else should. It refers to cats-in-the-software, if anything.

Section 4: Variations on the Vat

I've argued that the realist should concede that the brain's word "cat" does not refer to cats. And that claim, along with disquotation, is all that's needed for the BVA to be sound. The upshot of the argument so far is therefore this: The best strategy for the realist is not to attack the BVA itself. The BVA's assumptions about reference are too conservative for a plausible form of reference magnetism to reject them. That means that the realist must instead attack the anti-realist modus tollens that draws its support from the BVA. The relevant premise of that modus tollens was this: "Pervasive error is not a coherent possibility". The best course for the realist is to *accept* the BVA—to accept that the BIV scenario is self-undermining and thus to give up on the possibility that she is a brain in a vat—but to break the link between that conclusion and the incoherence of pervasive error. Before presenting what I think is the best strategy for accomplishing this, I consider one that may spring more quickly to mind.

The skeptical hypothesis Putnam asks us to consider is extremely specific. But there are a lot of skeptical hypotheses. I might be dreaming. I might be deceived by an evil demon. The world might have come into existence five minutes ago. I could even be a brain, but under circumstances different from those Putnam describes. There could be a programmer, or I could be recently envatted, or recently devatted, etc. Yet while Putnam's skeptical scenario is highly specific, the conclusion his BVA is meant to support—that pervasive error is incoherent—is extremely general. To show that radical falsehood of ideal theory is incoherent, it's not enough to show that *one specific* skeptical hypothesis is self-undermining. You would need to show that *all* skeptical hypotheses radical enough to do the trick are self-undermining. Why couldn't the realist simply concede the point about Putnam's BIV scenario, and then cheerfully point to any of the other skeptical hypotheses in the neighborhood to drive her wedge between ideal theory and truth?

I'm not entirely sure that some version of this response won't work. But if it does, it won't work nearly as easily as this description makes it sound.[32] Skeptical hypotheses vary from one another in many different respects. But one way in which they differ is in degree of radicalness. So consider an example of the type of skeptical hypothesis the realist might point to as an alternative to Putnam's BIV scenario: the hypothesis of recent envattment. Last night my brain was scooped from my head and I am now, undetectably, living in a simulation of the world I previously inhabited. If that scenario obtained, my word "vat" *would* refer to vats and my word "brain" *would* refer to brains. There's good reason therefore to think that a BVA-style argument would not apply to this skeptical hypothesis, and that we therefore cannot dismiss it as self-undermining.

As bizarre as the scenario is, however, I submit that it is also significantly less radical a skeptical hypothesis than the BIV scenario.[33] After all, if I was envatted last night (and nothing else is stipulated to differ from the actual world), then most of my beliefs are still true! I believe that I live in a world filled with medium sized dry goods like vats and computers and bottles. I believe that Lexington is in Kentucky, that whales are mammals, and that most persons are not envatted brains. All of those beliefs succeed in representing the world outside of the vat—that's how this scenario avoids a BVA-type argument, if it does—but they are also *true* beliefs.

Of course, some of my beliefs about my immediate environment are false on this hypothesis. I believe that I am sitting in front of a computer, in an office, using my two hands to type, and that my brain was not recently removed from my body and put in a vat. I'm wrong about all of those things, despite my seeming-evidence to the contrary. The scenario is plenty radical enough to be of interest to an epistemologist. But it's not at all obvious that this pattern of deception is radical enough to satisfy the metaphysical realist. Here's what that would sound like: "Because truth is non-epistemic, even our very best theory could turn out to be hopelessly, radically mistaken. After all, I might be a recently envatted brain with largely true beliefs about the world in general but lots of mistaken beliefs about what happened last night and what's going on in my immediate vicinity!" It doesn't sound quite as convincing.

The worry about the "alternative skeptical hypothesis" strategy is this: Think of the complete set of skeptical scenarios as partially ordered along a scale of *radicalness*. How to generate such a ranking would of course be a difficult question in itself. It might simply involve the number of our beliefs rendered false, were the hypothesis in question to obtain. More plausibly, it might privilege some of our beliefs as more central than others, and scenarios that target beliefs central to our overall theories as more

[32] This section owes a great deal to Button (2013), ch. 15. My presentation differs in important respects however, and my goals in making the argument are different.
[33] See Wright (1992), p. 87-88.

radical than scenarios that target more peripheral beliefs.[34] For my purposes, fortunately, a rough and ready understanding of the scale will do. At the high end of the scale are scenarios where we're deceived about everything or almost everything. At the low end of the scale are scenarios where we're deceived about only a very limited set of facts. Inhabitants of scenarios at the low end of the scale interact with the same kinds of things that we do, and can thus use their words to refer to the same things we do. Therefore we can't rule out, by virtue of describing those scenarios, that we are those speakers. Skeptical hypotheses at the low end of the scale really do evade the BVA. But of course they are also less skeptical.

Skeptical hypotheses at the high end of the scale are truly radical. Deception of the kind described in these hypotheses represents genuinely pervasive error. But of course these are also the scenarios where BVA-style arguments will most naturally apply. Speakers in scenarios like these—Putnam's BIV scenario, Descartes's evil demon scenario, Russell's five-minute hypothesis[35]—are so thoroughly deceived about their environment as to lack the capacity to describe it. So, in describing it, we demonstrate that we are not those speakers. The more radical a skeptical scenario is, the more likely that Putnam's argument will attach, and the more likely we can dismiss the scenario as self-undermining and stop worrying about it.

Somewhere in the middle, as you move up the scale, is the place where hypotheses change from coherent to self-undermining. Knowing where that threshold lies would tell us just how radically it is possible to be deceived, and thus just how false an ideal theory could be. Is the threshold far enough along the spectrum for the corresponding amount of error to count as *pervasive*? It's anything but clear. Button[36] argues that it is simply impossible to identify the point at which skeptical hypotheses become susceptible to the BVA. But even if we could locate the threshold, it would not be enough. We would also have to know whether that threshold is high *enough* for the realist. That is, we would have to say precisely how much potential for falsehood there must be for the realist to be satisfied that truth really is non-epistemic.

These tasks represent a tremendous amount of careful and difficult (and possibly—if Button is right—futile) philosophical work, jointly constituting a project as far as could be from simply shrugging one's shoulders and picking out another skeptical hypothesis. Perhaps, *pace* Button, the threshold for radicalness could be identified with precision and, *pace* my own hunch, some form of metaphysical realism deserving of the name could demonstrate itself to be content with skeptical scenarios lower in radicalness than that threshold. But conceding the point about Putnam's BIV scenario starts us on a slippery slope. And the general applicability and conservative nature of the BVA's premises make it hard to see how the brakes could be applied while the deception is still reasonably thorough. It's easy to

[34] Thanks to Mahrad Almotahari for helpful discussion on this point.
[35] Russell (1921).
[36] Button (2013), Ch. 15.

imagine getting argued far enough down the scale as to render this strategy pretty deeply unsatisfying to a serious realist.

Section 5: Ideology in the Vat

I submit that the realist should take a very different approach. Having signed up for the idea of objective metaphysical joints in nature, the realist is in a position to draw a distinction that matters greatly in this context. The distinction is between a theory's *truth* and its *overall epistemic success*. What the realist should insist, I maintain, is that it is not enough that a theory's claims be true. The theory should also employ concepts that cut the world close to the metaphysical joints. Two theories might both be made up entirely of true claims, but if theory A employs concepts that are metaphysically natural and theory B employs concepts that are gerrymandered and gruesome, then A is *better* than B. Crucially, it is not "better" in some instrumental sense. In virtue of employing concepts that cut the world closer to the natural joints, A is *epistemically* better: it more accurately characterizes what the world is like.

On the first page of his book-length defense of a Lewisian notion of structure, Sider asks us to imagine a world composed only of fluid, red on one side and blue on the other. The inhabitants of this world pay no mind to the plane dividing red from blue, however, but rather describe the world in terms of a different divide running diagonally across the fluid, each side incorporating a bit of red and a bit of blue. The inhabitants employ predicates corresponding to this diagonal plane, and they use those predicates—"bred" and "rue"—in making claims about various regions of their world. Sider's comment on the situation is worth quoting in full:

> It is almost irresistible to describe these people as making a mistake. But they're not making a mistake about where the red and blue regions are, since they make no claims about red or blue. And they make no mistakes when they apply their own concepts. The regions that they call "bred" are indeed bred, and the regions they call "rue" are indeed rue. The problem is that they've got the wrong concepts. They're carving the world up incorrectly. By failing to think in terms of the red/blue dividing plane, they are missing something. Although their beliefs are true, those beliefs do not match the world's structure.

Sider adopts from Quine the term *ideology* to denote the conceptual toolkit from which a theory is constructed. The people Sider describes have a theory that is epistemically unsuccessful. But it is not epistemically unsuccessful in virtue of being false. It is epistemically unsuccessful in virtue of employing a metaphysically non-natural ideology.

It's important at this point to be clear about the options available to the proponent of reference magnetism. Someone like Weatherson—and indeed many advocates of reference magnetism—might say that if the inhabitants of this world get *close enough* to the red/blue divide with their applications of "bred" and "rue," then the plane dividing red from blue will exert the "magnetism" of the view's title,

"attracting" the reference to the natural joint and making it the case that the speakers' terms pick out red and blue after all. Thus, when such a speaker points to a blue region of fluid and calls it "bred"—a term that on this view turns out to mean red—we can say that they are speaking falsely. We can thus describe the sense in which their theory is unsuccessful in the traditional way. The speakers make false claims about their world.[37]

Sider's remark makes it clear, however, that he thinks not all cases should be described in this way. As we saw in Section 3, the eligibility constraint is to be balanced against the other factors that play a role in reference determination. Whatever one's other views in semantics or metasemantics, it should be uncontroversial that speakers' usage of their terms is one crucial form of evidence about the meanings of those terms.[38] That evidence can be outweighed by other factors—evidence that certain patterns of usage should be explained pragmatically rather than semantically, for example, or a philosophically well motivated application of some form of semantic externalism. But there is no reason to think that any of those factors need apply in Sider's example. If the inhabitants of Sider's world are systematic enough in their dispositions to apply "bred" and "rue," if they do not tend to retract their claims when the red/blue divide is demonstrated to them, if the more charitable interpretation does a better job of playing an explanatory role in our overall theory of the speakers' thought and talk, then the right thing to say really is the more charitable thing: that "bred" picks out the bred region of their world and "rue" picks out the rue region.

Sider's point in the quoted passage is that to go this route is not to concede that these strange speakers have gotten their world *right*. For a theory to *succeed* epistemically, it must both (a) be true and (b) employ an ideology that cuts close to the joints. A theory can therefore *fail* epistemically in two independent respects. It can be false, or it can employ an ideology that does not cut close to the joints. Nothing prevents the realist from describing these speakers' true theory as an epistemic failure.[39]

---

[37] Note that the motivation for this kind of application of reference magnetism really has nothing to do with the model theoretic argument. If the MTA's perverse interpretations were still on the table, we could not even make sense of the notion of speakers' "getting close" to a natural joint, since to imagine that their usage has gotten them in the neighborhood of a joint is already to suppose that some interpretations are privileged over others. Whether this downstream application of the eligibility constraint can find independent motivation is, I think, an open question, given the considerations raised in this section. (See Sundell (2011) for discussion. In that essay, I rashly took myself to be objecting to reference magnetism in general. It would have been better to accept the view's application in addressing the MTA, to argue for the independence of the two types of applications, and to focus explicitly on cases where the view is used to motivate a choice among non-perverse interpretations, as it is in the whale/fish case.)

[38] See Plunkett and Sundell (2013), especially p. 16-17 and section 6.1.

[39] Burgess and Plunkett (2013a) and (2013b) give the label *conceptual ethics* to normative and evaluative questions about thought and talk. Evaluating scientific concepts with respect to their degree of naturalness is a perfect example of an issue in conceptual ethics. But metaphysical naturalness is not the only measure by which we might judge a concept. In other domains, we ask different things of our concepts, and even in science there may be a role for concepts that are held to standards other than naturalness.

Think then what a realist who has internalized this observation might say about Putnam's brain in a vat. Given highly plausible assumptions about reference, there is no way around the fact that the brain's best theory of its world is largely true. But as to whether the brain's theory is an epistemic success, that question remains open. And it turns out that the brain's largely true theory fails in crucial respects. For example, the brain lives in a world made up of electrons and quarks and brains and vats. But it has no words for any of these things. It lacks the conceptual resources to describe the world at the fundamental level of electrons or quarks, or even at the less fundamental level of ordinary physical objects. And the theory that it does have—the theory made up largely of true claims about bits of software—is constructed out of categories like electron-in-the-software and quark-in-the-software and cat-in-the-software. Those categories, while not irredeemably gruesome—you and I might make reference to them in a theory of the brain's mental states—do not come close to describing the world at the fundamental level. And yet, in the case of electron- and quark-in-the-software, they represent the brain's very best attempt at doing just that. Concede as much as you like about the truth of the brain's theory. The brain *is* radically deceived.

Now this observation by itself is no response to the BVA. As we saw in Section 2, the conclusion of the BVA is not: "Being a BIV wouldn't be so epistemically bad after all". The conclusion of the BVA is "I am not a brain in a vat." So pointing out one more way in which it would be bad, epistemically, to be a brain in a vat does nothing by itself to combat the argument. What it does accomplish, however, is to make salient the resources that the realist can draw on in mounting a response. The problem with Putnam's BIV-scenario came down to this: the BIV lacks the representational resources to describe its own predicament. So what the realist needs is a skeptical scenario that, unlike the BIV-scenario, could be represented by someone unfortunate enough to be in it.

So consider the following simple worry: "I could be in a situation preventing even my best theory from describing the world at the fundamental level." A few things to note about such a worry. First, the worry is *not* that I am not in fact sitting at a table, in an office, using my two hands to type a paper. In entertaining this worry I do not call into question the truth of any of those things. Second, please don't ask me to get more specific about the nature of the situation in question. What I've described is precisely the worry that I lack the representational resources necessary to characterize the world at a level fundamental enough to give such a description. Finally, and most importantly, observe that it is entirely possible for someone unfortunate enough to be in this predicament to nevertheless have the representational resources to entertain it.

To see this, consider again Putnam's BIV. Now, we know that *we're* not brains in a vat. If the above worry obtains for us, the relevant situation will have nothing to do with brains or with vats. Brains and vats are ordinary objects to which we successfully refer and about which we make largely true claims. If we are in such a predicament, it will have to do with objects and circumstances that our beyond our

representational capacities, just as brains and vats are beyond the BIV's. But that point can be set aside. We can look to the BIV as an example of the relevant type of unfortunate soul and ask about the kinds of worry it can entertain. The brain lacks the representational resources to describe its world at a fundamental level. So: can it wonder whether that's true? What are the representational resources necessary to have this worry? Well, you would need to be able to refer to *theories*. And you would need to be able to think about *ideology*. You would need to have thoughts about *situations* and *descriptions*. And, crucially, you would need a notion of *relative fundamentality* and the ability to worry that your own theory employs an ideology *less* fundamental than a *maximally fundamental*, fully epistemically successful theory would employ.

There is, I submit, no good reason to think that the BIV lacks any of these things. There's not even *prima facie* reason to doubt that the brain could refer to *theories*, *situations*, and the like. The harder question is whether it can share with us a notion of *relative* and of *maximal fundamentality*. It can. After all, the brain can contrast its word "green" and its word "grue". It can categorize objects by whether they are "cats" or "cwats". It can represent and contrast the plus function and the quus function. That the brain's word "green" doesn't pick out the color green, but rather green-in-the-software is irrelevant. That the more natural property of being a cat-in-the-software is not, all things considered, terribly natural is irrelevant. You don't need to represent properties across the complete spectrum of naturalness in order to have a notion of *relative* naturalness. And a notion of relative naturalness is enough to conceive of something's being maximally natural. Compare: You don't need to have come in contact with absolute zero in order to formulate questions about its nature or existence. And you certainly don't need to have had contact with the full range of temperatures in order to categorize the weather around here in terms of relative coldness.[40] When we use the phrase "colder than," we pick out the same relation as speakers in a world where everything is closer to absolute zero than anything you or I will ever experience. Unlike truth and falsity, naturalness is a matter of degree. Being *somewhere* on the spectrum is enough to provide the conceptual resources necessary to wonder just how far you are from the end point.

This response is similar in some ways to responses in Smith (1984), Wright (1992), Forbes (1995), and Chalmers (1995). Those authors each advance some version of an "exemplification" response to the BVA, wherein we concede to Putnam that we are not in Putnam's BIV scenario, but maintain that we might nevertheless be in some more schematically characterized situation "relevantly similar" to Putnam's BIV. Davies (1997) argues forcefully that such responses cannot work. Responding to Smith's formulation, Davies considers the suggestion that the BIV stands in the "delusive relation" to its world, and that while we cannot worry that we are BIV's, we can worry that we stand in that same delusive

[40] Thanks to Josh Brown for helpful discussion on this point.

relation to ours. For that to work, however, it must be the case that the BIV's phrase "delusive relation" refers to the same relation as our phrase "delusive relation". Otherwise, again, the resident of the "delusive scenario" would be incapable of representing its own predicament, and so, likewise, our phrase "delusive relation" would pick out a relation that we could not stand in to our own world.

Using Putnam's convention of capitalized "WORLD" to denote the objective world posited by the metaphysical realist, Davies notes that

> a BIV's ability to use the label 'the delusive relation' to refer to a specific real-WORLD relation between real brains and other things in THE WORLD seems no less problematic than its ability to use the word 'brain' to refer to real brains.And, to the extent that we are envisaging the possibility that our own epistemic situation is analogous to that of a BIV, we should not ascribe to ourselves the sorts of referential capacities that a BIV would lack.[41]

What does the BIV's phrase "delusive relation" refer to? The BIV gives that phrase meaning by pointing to instances in its own experience. But those instances—brains-in-the-software "in" vats-in-the-software, for example—exemplify a relation, not between real brains and an external world, but between various images or bits of software. That's nothing like the actual relation the BIV stands in to the world outside its vat. Similarly, our phrase "delusive relation"—exemplified by pointing to brains in vats—does no better a job of representing a predicament we might actually be in than our phrase "brain in a vat".

Despite its similarities to other exemplification responses, however, the response I have described is not subject to this objection. The relation between a BIV-in-the-software and the "world" outside the vat-in-the-software is nothing like the relation between the BIV itself and the world outside its vat. That's why the BIV and we mean different things by "delusive relation". However, in drawing comparisons not between epistemic situations but rather between concepts, there is no parallel reason for worry. The relation between the BIV's less natural concepts and its more natural concepts is the same relation that stands between our own less natural concepts and our own more natural concepts. The concepts CATS-IN-THE-SOFTWARE and CWATS-IN-THE-SOFTWARE are both farther down the naturalness scale than the concepts CATS and CWATS. But the *difference* between the respective members of each pair is the same. It is a difference in fundamentality.[42] The phrase "stands in the delusive relation to" may indeed pick out different relations depending on whether it is a phrase in English or vat-English. But there is no analogous reason to think that we and the BIV do not mean the same thing by "is less natural than".

---

[41] Davies (1997), p. 53.

[42] If the fact that both cats-in-the-software and cwats-in-the-software are bits of software makes you nervous, the comparison of "plus" and "quus"—which pick out the same pair of functions inside and outside of the vat—suffices to make this point.

So, the brain in a vat is in a situation preventing it from carving its world at the joints. And the brain in a vat—while it cannot worry that it is a brain in a vat—*can* worry that it is in a situation preventing it from carving its world at the joints. And since naturalness of ideology is one component of epistemic success, this worry is very much a *skeptical* worry. The hypothesis that our concepts—though we may deploy them correctly—are massively more gruesome than we'd hoped or thought is a skeptical hypothesis every bit as nightmarish as the Brain in a Vat or the Evil Demon or the Young World or any of the others. Indeed, it's worse than any of them, because it describes a pervasive form of deception that sidesteps entirely the fine-grained analysis required to determine whether it is sufficiently *non-radical* to evade the BVA. The most radical of the traditional skeptical hypotheses can be dismissed as self-undermining. The less radical, while they cannot be dismissed—and thus raise interesting epistemological questions of their own—have little to teach us about the prospects for realism. The worry about the nature of our concepts suffers from neither of those problems.

So where exactly is the mistake in the BVA-powered anti-realist modus tollens? The mistake is in focusing exclusively on truth. For a certain type of non-realist, it might well be that truth is all there is to epistemic success. If you don't believe in joints, then you won't go in for the idea of evaluating concepts by how close they get to them. It thus might come to appear as if the realist must drive a wedge between ideal theory and *truth*, in particular, if she is to demonstrate that the fundamental nature of the world is independent of our thought and our inquiry. But what Putnam is in a position to demand, and what the realist needs, is not a wedge between ideal theory and truth. What the realist needs is a wedge between ideal theory and *epistemic success.* And not everyone agrees about what epistemic success involves. For the realist—supposing she adopts the strategy I advocate—a theory of the natural world aims to succeed along two dimensions, and thus it risks failure along both of those dimensions. Both of those dimensions are epistemic, and thus failure along either dimension counts as *error* in any relevant sense. Plausible constraints on meaning and reference may demonstrate that an ideal theory is guaranteed to be largely true.[43] Even so, an ideal theory could be hopelessly, radically mistaken. It could be constructed out of a hopelessly, radically non-fundamental ideology. In other words, the BVA may demonstrate that pervasive *falsehood* is incoherent, but that is simply too specific to refute metaphysical realism. What would refute metaphysical realism is the incoherence of pervasive *error*, and sadly enough, there is nothing self-undermining about that.

Conclusion

[43] They do not guarantee even an ideal theory to be *entirely* true. The less radical, BVA-resistant skeptical scenarios are enough to show that.

Perhaps the strategy I have described will still strike a realist as overly concessive. Although I claim to have cut off the anti-realist modus tollens, I've suggested that the realist concede the both the soundness and the wide applicability of the BVA, and thus concede that ideal theory is in fact guaranteed to be largely true. But doesn't that just amount to a concession that truth itself is epistemic? Not at all. Consider the question of *why*, given the considerations raised above, an ideal theory is guaranteed to be largely true. Whether the claims we make are true depends on what it is we've claimed. What it is we've claimed depends on what our words mean. What our words mean may not depend *entirely* on us. Under the supposition that reference magnetism is true, for example, there are other factors involved. But what our words mean depends *largely* on us. A semantic theory that paid no mind to speakers' usage, intuitions, dispositions, or surroundings would not be a plausible theory of language. Precisely how close a semantic theory must hew to speakers' usage or first-order intuitions is up for debate. But the fact that a semantic theory must display some sensitivity to our usage, dispositions, surroundings, causal relationships, and so on is not up for debate, as a fan of the BVA will be the first to admit.

That sensitivity by itself is enough to guarantee the truth of much of what we say. But note that we've said nothing so far about the theory's being *ideal*. The considerations raised above demonstrate that *any* theory of ours, ideal or not, is guaranteed to be largely true, simply in virtue of being comprised of statements in our language. The lesson of the BVA is thus not that truth is epistemic. The lesson is simply this: *Truth is semantic*. The fact that any theory of ours is guaranteed to be largely true is a function of the relationship between truth and meaning, and between meaning and use. By itself such a guarantee has none of the striking metaphysical consequences it might at first appear to. Of course, if truth were all there was to epistemic success, then the guaranteed truth of our theories would indeed be difficult to integrate into a metaphysical picture on which the nature of the world is fully independent of us and our inquiry. It would be hard to see how, given such a guarantee, we could get things seriously wrong. And we can definitely get things seriously wrong. But truth is not all there is to epistemic success. We don't just have to figure out what to say—we also have to select the words and concepts with which we say it. And in our choices about which words and concepts we should be using in inquiry, there are no guarantees of success. We have to hope for the best and keep inquiring.

Works Cited

Brueckner, Anthony L. “Brains in a Vat.” *Journal of Philosophy* 83.3 (1986): 148-167. Print.

Brueckner, Anthony L. “If I Am a Brain in a Vat, Then I Am Not a Brain in a Vat.” *Mind* 101.401 (1992), pp. 123-8.

Burgess, Alexis and David Plunkett. “Conceptual Ethics I.” *Philosophy Compass* 8:12 (Dec 2013, A): 1091-1101. Print.

Burgess, Alexis and David Plunkett. “Conceptual Ethics II.” *Philosophy Compass* 8:12 (Dec 2013, B): 1102-1110. Print.

Button, Tim. *The Limits of Realism*. Oxford: Oxford University Press, 2013. Print.

Button, Tim. “Brains in Vats and Model Theory.” *The Limits of Realism*. Oxford: Oxford University Press, 2013. Print.

Chalmers, David J. “The Matrix as Metaphysics.” *Philosophers Explore the Matrix*. Ed.: Christopher Grau. New York: Oxford University Press, 2005. 132-177. Print.

Davies, D. “Putnam’s Brain Teaser.” *Canadian Journal of Philosophy* 25.2 (Jun 1995): 203-228. Print.

Davies, D. “Why One Shouldn’t Make an Example of a Brain in a Vat.” *Analysis* 57.1 (1997): 51-59. Print.

DeRose, K. "How Can We Know That We're Not Brains in Vats?", *The Southern Journal of Philosophy* 38 (2000), Spindel Conference Supplement: 121-148.

Ebbs, Gary. “Realism and Rational Inquiry.” *Philosophical Topics* 20.1 (Spring 1992): 1-34.

Fodor, Jerry A. *Psychosemantics*. Cambridge, MA: MIT Press, 1987.

Forbes, G. “Realism and scepticism: brains in a vat revisited.” *The Journal of Philosophy* 92 (1995): 205-22.

Lewis, David. “New Work for a Theory of Universals.” *Australasian Journal of Philosophy* 61.4 (1983): 343–77. Print.

Lewis, David. “Putnam’s Paradox.” *Australasian Journal of Philosophy* 62.3 (1984): 221-236. Print.

Millikan, Ruth. *Language, Thought, and Other Biological Categories*. Cambridge: MIT Press, 1987. Print.

Plunkett, David and Tim Sundell. “Disagreement and the Semantics of Normative and Evaluative Terms.” *Philosopher’s Imprint* 13.23 (Dec 2013): 1-37. Print.

Putnam, Hilary. “Realism and Reason.” *Proceedings and Addresses of the American Philosophical Association* 50.6 (Aug 1977): 483-498. Print.

Putnam, Hilary. “Models and Reality.” *The Journal of Symbolic Logic* 45.3 (Sep 1980): 464-482. Print.

Putnam, Hilary. *Reason, Truth, and History*. Cambridge: Cambridge University Press, 1981.

Putnam, Hilary. “Replies.” *Philosophical Topics* 20.1 (Spring 1992): 247-408. Print.

Russell, Bertrand. *The Analysis of Mind*. London: George Allen & Unwin Ltd: 1921.

Sider, Theodore. *Writing the Book of the World*. Oxford: Oxford University Press, 2012. Print.

Smith, “Could We Be Brains in a Vat?”, *Canadian Journal of Philosophy*, 14.1 (1984): 115–123.

Sundell, Timothy. (2011). “Disagreement, Error, and an Alternative to Reference Magnetism.” *Australasian Journal of Philosophy* 90.4 (2011): 743-759. Print.

Tymockzko, T. “In Defense of Putnam’s Brains.” *Philosophical Studies,* 57.3 (1989): 281-297. Print.

Warfield, T. A. “A Priori Knowledge of the World: Knowing the world by knowing our minds”. *Philosophical Studies* 92.1/2 (1998). pp. 127-47.

Weatherson, Brian. “What Good Are Counterexamples?” *Philosophical Studies* 115.1 (2003): 1-31. Print.

Williams, J. Robert G. “Eligibility and Inscrutability.” *Philosophical Review* 116.3 (2007): 361-399. Print.

Wright, Crispin. “On Putnam’s Proof that We Are Not Brains in a Vat.” In P. Clark and B. Hale (eds.), *Reading Putnam*. Cambridge: Blackwell, 1992. Pp. 216-241. Print.