META: RESEARCH IN HERMENEUTICS, PHENOMENOLOGY, AND PRACTICAL PHILOSOPHY
VOL. X, NO. 1 / JUNE 2018: 149-159, ISSN 2067-3655, www.metajournal.org

Putnam's Argument that the Claim that We are Brains-in-a-vat is Self-Refuting

Richard McDonough
Arium Academy of Arts and Sciences, Singapore

Abstract

In *Reason, Truth and History*, Putnam provides an influential argument for the materialist view that the supposition that we are all "actually" brains in a vat [BIV's] is "necessarily false". Putnam admits that his argument, inspired by insights in Wittgenstein's later views, is "unusual", but he is certain that it is a correct. He argues that the claim that we are BIV's is *self-refuting* because, if we *actually* are BIV's, then we cannot refer to real physical things like vats. Although the present author agrees, fundamentally, with Heidegger's view that we are essentially "in a world" (a real world, not a private "vat-world"), and, therefore, with Putnam's *conclusion* that we cannot possibly be BIV's, the paper argues that Putnam's *argument* is fallacious. The proper conclusion to draw from Putnam's argument is that asserting that one is a BIV is beyond the limits of a BIV's (private) language. That is, Putnam only shows that if we actually are BIV's, then we cannot think or assert either that we *are* or that we *are not* BIV's. It does not show that we are not "actually" BIV's. The cogency of this criticism is illustrated with a concrete "science-fiction" example.

Keywords: Putnam, Heidegger, Wittgenstein, brains in a vat, intrinsic reference, private language

> "Being in a world is something that belongs essentially … to Dasein" (Heidegger, *Being and Time*, § 4).

In *Reason, Truth and History* (*RTH*), Putnam provides an influential argument for the view that "the supposition that we are actually brains in a vat [BIV's] … *cannot possibly be true* [P's emphasis]" (*RTH*, 7). By the thesis that we are BIV's he does not mean that some people some of the time are BIV's, but that "everything [in human experience] is a collective hallucination" produced either by an "evil scientist" or an "absurd" universe consisting of machines tending to BIV's (*RTH*, 6-7, 130-131). Putnam regards this view, which he attributes, somewhat tongue in cheek, to most Australians, but especially to "the Guru of Sydney" (who he names "David"), as "incoherent" (*RTH*, 131).[1] Putnam admits that his argument, broadly inspired by insights in Wittgenstein's later views, is "unusual", but he is certain that "it is a correct argument". (*RTH*, 7) He argues that David's assertion that we are brains in vats (hereafter WBV) is *self-refuting* because, if we *actually* are BIV's (hereafter WAABV's), we cannot refer to real physical things like vats.[2] (*RTH*, 7) BIV's "cannot think or say that they are brains in a vat". (*RTH*, 14) Thus, WBV is "necessarily false". (*RTH*, 7-8, 15) However, Putnam admits that WBV "violates no physical laws and is perfectly consistent with everything we have experienced".[3] (*RTH*, 7) But then how can one conceivably demonstrate that WBV cannot *possibly* be true? Putnam explains that one must not take "physical possibility as the touchstone of what might actually be the case", a common mistake in misguided ages that "takes physics as [their] metaphysics". (*RTH*, 15) The present paper argues that Putnam's argument is fallacious. The fact, if it is one, that WBV is self-refuting on Putnam's grounds is logically irrelevant to the question whether we "actually" are BIV's. The proper conclusion to draw from Putnam's argument is that asserting that one is a BIV is beyond the limits of a BIV's (private) language. That is, Putnam only shows that if we actually are BIV's, then we cannot think or assert either that we *are* or that we *are not* BIV's. But it does not show that we are not "actually" BIV's. It is worth emphasizing that although the present author agrees fundamentally with Heidegger's view in *Being and Time* that human beings are essentially "in-the-world", and, therefore, agrees with Putnam's *conclusion* that the view that we are all BIV's is incoherent, the problem is that Putnam's *argument* does not eliminate the possibility that we "actually" are BIV's.

Putnam lists two main assumptions of his argument. The first is that "magical theories of reference", which hold that representations can intrinsically refer to something beyond themselves, are wrong. (*RTH*, 16-17, 21).[4] The second is that one can refer to a given kind of things, say trees, only if one has some causal interaction with them, or with things in terms of which they can be described. (*RTH*, 16-17)

The first premise is needed because, if brains contain representations which intrinsically refer to things beyond themselves then there is no reason why BIV's could not think or assert that they are BIV's. Although philosopher's do not normally, in public at least, declare support for magical theories of reference, the point needs to be stated because it is alleged that many philosophers, presumably without explicitly recognizing this, *tacitly* assume that there are representations that intrinsically refer to things outside themselves, e.g., Fodor, Chomsky and others hold that mental representations display their meaning intrinsically" in a way that no mere physical sign can do (Goldberg 1983, 196-210).

The second premise is needed because causal interaction with things in the external world is precisely what BIV's lack. Putnam does not specify precisely how much causal interaction with objects in the world is required if one can refer to them but he insists that a certain level of causal interaction is necessary. For example, if a mad scientist produces images of trees for BIV's by using only electronic impulses without the causal involvement of trees, and if a BIV's entire experience is produced in this way, then that BIV cannot refer to trees.

Deprived of magical representations and causal interactions with trees or things connected with trees, BIV's can only think, not in English, which refers to trees, but in vat-English,

> The truth-conditions for 'There is a tree in front of me' … in vat-English are simply that the tree in the image be 'in front of' the 'me' in question – in the image – or, perhaps, that the kind of electronic impulse that normally produces this experience be coming from the automatic machinery, or, perhaps that the feature of the machinery that is supposed to produce the 'tree in front of one' experience be operating. (RTH, 14)

Since BIV's are trapped inside their envatted experience (or, perhaps, being generous, the impulses and machinery that produce that envatted experience), their sphere of reference is severely limited. It is this that Putnam claims makes a BIV's assertion that it is a BIV self-refuting.

Using the acronyms, Putnam's central claim that if we actually are BIV's, then "We are BIV's" is self-refuting becomes: "If WAABV, then WBV is self-refuting". Call this Putnam's Conditional or PC! WAABV is the antecedent. "WBV is self-refuting" is the consequent. But why is WBV self-refuting? "If … we really are brains in a vat, then what we now mean by 'we are brains in vats' is that *we are brains in a vat in the image* or something of that kind (if we mean anything at all) [P's emphasis]" (*RTH*, 15). That is, if we "actually" are BIV's, then WBV is stated in vat-English, not English proper. Since vat-English cannot refer to real physical vats, WBV cannot refer to real physical vats and, therefore, cannot assert that we are "actually" are BIV's. Putnam does *not* claim that WBV is self-refuting because it *contradicts* itself. WBV is self-refuting because if we actually are BIV's then our attempt to assert that we are BIV's necessarily fails.

Putnam is vague about what WBV can mean. He lists three possibilities, 1.) It means we are brains in a vat "in the image", 2.) It means "something of that kind", 3.) It may not mean anything at all. (*RTH*, 15) Putnam is not admitting to some flaw in his formulation. The implicit claim is that there is an objective unclarity in WBV and Putnam is merely reflecting that unclarity. However, it is this that first signals that Putnam's argument is fallacious.

Consider the third of Putnam's glosses on WBV, that it may not mean anything at all. But if it does not mean anything, then it is neither true nor false, in which case it cannot "refute" anything in the sense of showing that it is false. Since this is one of Putnam's possible readings of WBV, WBV is not "self-refuting" in the sense required to show that WBV is false, let alone "necessarily false". Indeed, if WBV may not mean anything at all the category of "self-refutation" is misapplied here. In fact, much of the obscurity in Putnam's argument traces to his odd notion of self-refutation. However, one might think that this objection is unfair. Perhaps Putnam is in that third gloss merely registering puzzlement about what WBV can mean while his first and second readings of WBV are meant more seriously.

Consider Putnam's suggestion that WBV means that 1.) we are brains in a vat "in the image" or 2.) "something of that kind". Putnam's second reading is given because it is so hard to know how, precisely, to specify WBV's meaning. Does WBV mean that in a BIV's *image* of its world we are BIV's, or that in the BIV's image of the world we are images of a brain in an image of a vat—or something else? In any case, WBV's meaning must, if we actually are BIV's, be specified in vat-English, not English. Thus, WBV's statement that we are BIV's cannot mean "we" in the sense in which this is understood in English to refer to real human beings "in the world". It can only mean "we" in the sense available in vat-English. The "we" in WBV has to mean the "we" in a BIV's private vat-language, roughly, whatever concept vat-English can have of the human family. The same holds for the meanings of "brain" and "vat" in vat-English. In order to distinguish these words in vat-English from their counterparts in normal English one can subscript them, "brain$_{ve}$", "vat$_{ve}$", etc. The words in normal English are not subscripted. Thus, one can replace WBV by WBV$_{ve}$: "We$_{ve}$ are brains$_{ve}$ in vats$_{ve}$". Plugging in the subscripted terms, PC becomes: "If WAABV, then WBV$_{ve}$ is self-refuting". The appearance that WBV$_{ve}$ is self-refuting (or necessarily false) arises because it attempts, but necessarily fails, to affirm what is plainly affirmed in the antecedent, WAABV (that we "actually" are BIV's). However, WBV$_{ve}$ is not necessarily false in the case envisaged by Putnam. The two assertions, WAABV and WBV$_{ve}$, written in different, but superficially similar looking languages, English and vat-English, are about entirely different things. WAABV is about physical human beings "in the world" and about brains and vats described in everyday English. WBV$_{ve}$ is, roughly, about the brains and vats "in the image" ("or something like that") as described in vat-English. There seems no obvious reason why one cannot provide truth-conditions for WBV$_{ve}$ ("We$_{ve}$ are BIV$_{ve}$'s") that are independent of the truth-conditions for the normal English WAABV in the antecedent, "We actually are BIV's". The illusion that there is a conflict between WBV$_{ve}$ and WAABV, thereby making WBV$_{ve}$

necessarily false, is fostered by the false assumption that the words in WAABV and WBV$_{ve}$ are about the same things. It is, admittedly, difficult to know how to fill out the truth conditions for WBV$_{ve}$, as Putnam acknowledges, but that just emphasizes the point. To the degree that it is difficult to fill out the truth conditions for WBV$_{ve}$ (in the consequent), then it is difficult to assert that they are "refuted" by the truth conditions for WAABV (in the antecedent). There is, therefore, no obvious reason (or none given so far by Putnam) why we cannot actually be BIV's and yet the vat-English "We$_{ve}$ are BIV$_{ve}$'s" is also true. A science fiction example to illustrate this point is provided later.

This highlights the self-admitted "unusual" logic of Putnam's argument. Putnam's key claim is that WBV$_{ve}$ is "self-refuting". But this is misleading. The reason WBV$_{ve}$ is supposed to be self-refuting is that since WBV$_{ve}$ is in vat-English, not English, WBV$_{ve}$ necessarily fails to state what it purports to state. However, the fact that WBV$_{ve}$ necessarily fails to state what it purports to state has *nothing* to do with the question whether we *actually* are BIV's. Since WBV$_{ve}$ is a statement in vat-English, the proper conclusion to draw from Putnam's argument is not that the statement in English that we are BIV's is necessarily false, but that a BIV, speaking vat-English, is not in a position to meaningfully assert *either* that we are or that we are not BIV's. Indeed, if we actually are BIV's (WAABV), then Putnam can no more deny that we are BIV's in his vat-English than the guru of Sydney, David, can affirm it in their shared vat-language. Putnam has demonstrated that the question whether one is a BIV is beyond the limits of a BIV's private vat-English. That is, he has demonstrated that if we actually are BIV's then neither Putnam nor David can even represent the thesis that we are BIV's in their private vat-languages. But that has *nothing* to do with the truth or falsity of the assertion, in normal English, that we are BIV's.

In order to illustrate this, consider the following science fiction example. Suppose that the what we call the human race is a vast science experiment by Arcturian scientists to study BIV-functioning. There is no planet called "earth" peopled by philosophers such as Putnam and David. Since the earthly

Putnam and David do not exist, call these Arcturian BIV's PutnamA and DavidA! PutnamA's and DavidA's beliefs that they belong to a distinguished history of earthly human philosophers has been induced by artificial electronic means by Arcturian scientists. Putnam cannot object to the example because he admits that this is physically possible and consistent with everything we have experienced.

Two of these Arcturian scientists, Trin and Tran, stand on a raised walkway surveying rows of experimental BIV's. Examining some computer printouts of BIV activity, Trin remarks: "Poor PutnamA over there in vat # 29 takes himself to have demonstrated that he and DavidA in the adjacent vat # 30 are not BIV's on the grounds that DavidA's claim that they are BIV's is self-refuting, when, as you and I can both see [with our electromagnetic sensory organs], they both manifestly are BIV's. We Arcturians easily grasp this because we. being creatures who live "in a world", speak a genuine worldly language (Arcturian-English). PutnamA's "refutation" of DavidA's claim that they are BIV's is only expressed in vat-English and that has nothing to do with the question whether he and David *actually* are BIV's. Unfortunately, the only way we could explain this to PutnamA is by showing him that he is only speaking vat-English, rather than English proper, but, deprived as he is of any contact with the real external (to his vat) world, he is not even in a position to grasp that he is speaking vat-English rather than English". Tran replies, "Well I do not deny that PutnamA is in a bad way, but it's not his fault. One cannot expect too much of a BIV. But DavidA is certainly no better off. One might think that DavidA is closer to the truth because he says that he and PutnamA are BIV's and they are, but that's an illusion. For DavidA takes himself to believe that he is a BIV when, in fact, confined to vat-English as the poor brain is, he is only capable of thinking that heve is a BIVve (the semantics for which even we Arcturian "beings in a world" cannot quite figure out). Neither one is any closer to the truth than the other. What we have here in PutnamA and DavidA are two BIV's, one of them arguing that they are BIV's and one of them arguing that they are not, and yet neither one of them is right because neither of them is even able to represent in their limited vat-language the thesis in English that they actually are BIV's, which is supposed to be at issue. Trin remarks, "I get that. But surely you admit that Putnam$_A$'s argument appears to show that David$_A$'s claim that they are BIV's is self-refuting. It fooled me". Tran replies, "Putnam$_A$ has shown that if they actually are BIV's, then David$_A$'s claim (WBV) cannot state what it purports to state (that they are BIV's). One can call that "self-refuting" if one wants, but the only thing that is "refuted" here is that WBV is the statement in English that it purports to be. Putnam$_A$ has shown that if we actually are BIV's, then David$_A$'s thesis is really only a claim in vat-English, not the claim in English that David$_A$ purports to assert and Putnam$_A$ purports to deny". Trin sums it up, "What Putnam$_A$ has really shown is that if they really are BIV's, then they do not even know, and *necessarily* cannot know, what language they are speaking?". "That's it", Tran replies, "a BIV's claim that they are thinking in a genuine worldly language" cannot be true". Trin cannot resist a final quip: "No doubt Putnam$_A$ and David$_A$ are in a bad way. But do not even get me started on Heidegger$_A$ in vat # 47, who has been vehemently insisting that he is essentially 'in-a-world' when we both see clearly that he is really only 'in-a-vat'".

As stated earlier, the present author agrees with Putnam (and, presumably, Heidegger) that the view that all human beings might be BIV's is incoherent. The problem is that Putnam's argument does not demonstrate that we, and Heidegger, are not in the peculiar envatted position represented by the above science fiction example. That is, Heidegger, the real earthly Heidegger "in-the-world", is correct that we are essentially "in a world", and Putnam may take himself to have demonstrated this, but his argument fails to do so. A new, and superior, argument is required to show that the view that we are BIV's is incoherent.[5]

In summary, Putnam's argument that we are not BIV's is fallacious. Putnam's PC boils down to this: If we actually are BIV's, then, since our language is vat-English rather than English proper, our attempt to assert that we are BIV's necessarily fails because BIV's cannot even represent the thesis in English proper (English "in the world") that we are BIV's. It can at most represent the thesis that we$_{ve}$ are BIV$_{ve}$'s. But, as the Arcturian example illustrates, PC is logically irrelevant to the truth of the ordinary English assertion that we "actually" are BIV's. Putnam intuition that if we actually are BIV's there is *something* very peculiar about David's WBV is correct. But what is wrong with WBV is not that it is necessarily false. What is wrong with it is rather that it necessarily cannot be the English assertion it purports to be. Putnam has only found a complicated way of saying that since a BIV-language is a private language, BIV's necessarily cannot either assert or deny in (public) English "in the word" that they are BIV's. That should not be surprising, and it has nothing to do with the *substantive* question whether we really are BIV's, as Trin and Tran can clearly see from their walkway above the vats containing the ever squabbling (or, more precisely, squabbling$_{ve}$) Putnam$_{A}$ and David$_{A}$. Heidegger is right that we are essentially "in-a-world", not "in a vat", but Putnam has not demonstrated that the claim that we are all "in a vat" is incoherent.

NOTES

[1] For a survey of the various responses to Putnam's argument, see S. Goldberg (2016). Goldberg's Introduction to the volume is quite useful. Putnam's "David" is, presumably, David Armstrong. Mumford (2007, 130) states that Armstrong's (2002) *A Materialist Theory of Mind* remains "an authoritative [and seminal] statement of Australian Materialism. See also Armstrong (1985)!

[2] One might be tempted to use the same acronym in the antecedent and consequent as follow: If WBV then WBV is self-refuting. However, Putnam distinguishes the antecedent and consequent by describing the antecedent as the thesis that we "actually" are BIV's. This is represented here by "WAABV." The limited similarity between the two theses is reflected in the limited similarity in the two acronyms. Indeed, the partial similarity between the two verbal formulations may help to facilitate Putnam's mistaken inference.

[3] Putnam only refers here to what we "have experienced," leaving open the possibility that something in our future experience might disconfirm the thesis, e.g., the mad scientist tending the vats reveals the BIV's real situation to them somehow. However, this is not directly relevant to the present paper.

[4] Putnam should not say "magical theories of reference," but, rather, "theories of magical reference". For it is not the theory per se that is magical but the kind of intrinsic reference envisaged by the theory that is allegedly magical.

However, this nuance does not affect Putnam's argument or the argument of the present paper.

[5] The present author holds that Wittgenstein's (2010, para's 243-271) "private language argument," may provide just such an argument, for BIV is just a physicalistic version of the view that human beings have a private language, but a discussion of that the private language argument is beyond the limits of this paper.

REFERENCES

Armstrong, David. 1968. *A Materialist Theory of Mind*. London: Routledge and Kegan Paul. [Routledge produced a revised 2nd edition in 2002.]

Armstrong, David. 1985. "Consciousness and Causality." In *Consciousness and Causality*, by David Armstrong and Norman Malcolm, 103-191 and 205-217. Oxford: Blackwell.

Heidegger, Martin. 1962. *Being and Time*. Translated by John Macquarrie & Edward Robinson. San Francisco: Harper.

Putnam, Hilary. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press. [Cited in text as RTH].

Goldberg, Bruce. 1983. "Mechanism and Meaning". In *Knowledge and Mind: Philosophical Essays*, edited by Carl Ginet and Sydney Shoemaker, 191-210. Oxford: Oxford University Press.

Goldberg, Sanford. 2016. *The Brain in a Vat*. Cambridge: Cambridge University Press.

Wittgenstein, Ludwig. 2010. *Philosophical Investigations*. Translated by Elizabeth Anscombe, P.M.S. Hacker, and Joachim Schulte. Hoboken: John Wiley and Sons.

Richard McDonough is an adjunct professor at the *Arium Academy of Arts and Sciences* in Singapore. His research interests are: Philosophy of mind, philosophy of language, history of philosophy (especially ancient Greek philosophy, Taoism, German Idealism, 20th-21st century continental philosophy), Wittgenstein, history of psychology, philosophical literature.
See his publications at https://philpapers.org/s/richard%20mcdonough
Current projects: A book on the development of Wittgenstein's philosophy from his *Tractatus* to his later writings, beginning with the *Philosophical*

Investigations. The *provisional* title of the book is *The Truth of the Argument of the Tractatus.* The book attempts to lay out the dialectical development of Wittgenstein's views of the *Tractatus* into his later views, beginning with the *Philosophical Investigations*, and argues for the superiority of the later views.

Address:
Richard McDonough
Arium Academy of Arts and Sciencees
35 Selegie Road
Parklane #09-28
Republic of Singapore
Email: wittgensteins_poker2007@yahoo.com