Published in *Studies in History and Philosophy of Science* (2004) 35: 227–236.
doi:10.1016/j.shpsc.2004.03.007

mark.sprevak@ed.ac.uk

Magic, semantics, and Putnam's vat brains

Mark Sprevak
University of Edinburgh

Christina McLeish
University of Cambridge

13 March 2004

In this paper we offer an exegesis of Hilary Putnam's classic argument against the brain-in-a-vat hypothesis offered in his *Reason, Truth and History* (1981). In it, Putnam argues that we cannot be brains in a vat because the semantics of the situation make it incoherent for anyone to wonder whether they are a brain a vat. Putnam's argument is that in order for 'I am a brain in a vat' to be true, the person uttering it would have to be able to refer successfully to those things: the vat, and the envatted brain. Putnam thinks that reference can't be secured without relevant kinds of causal relations, which, if envatted, the brain would lack, and so, it fails to be able to meaningfully utter 'I am a brain in a vat'. We consider the implications of Putnam's arguments for the traditional sceptic. In conclusion, we discuss the role of Putnam's arguments against the brain in a vat hypothesis in his larger defense of his own internal realism against metaphysical realism.

1 Introduction

Consider the possibility, familiar to us in many guises, that instead of being a living breathing human, you are a brain in a vat. Almost everything that you believe about the external world is false. Your body, your friends, your family, your home—none of these things exist. The earth, the sun, and the stars do not exist. The page that you are reading now doesn't exist. The world you live in is radically different from what your experiences have led you to believe. This world contains, amongst other things, a laboratory in which you in the form of your brain are kept alive. An evil scientist

runs the laboratory and electrically stimulates your brain so as to give you the illusion of normal life. You experience all the ups and downs of life: walking, talking, interacting with other humans, but none of this is real; none of it corresponds to anything we normally assume is 'out there' in the world.

How do you know that you are not in this situation now? If the evil scientist doesn't slip up (and we can assume he doesn't), his presence will be impossible to detect. If this is so, it would seem impossible for you to rule out that you are now, in fact, a brain in a vat. It is true that it does not seem to you or me now that we could be a brain in a vat—the world seems so real and immediate. But what specifically about your experience of your life tells you that you are not floating in a vat of nutrients? The evil scientist is responsible for how real everything appears—he is even responsible for how absurd the scenario sounds. Perhaps we might want to say that such a thing is *possible*, but highly unlikely. On reflection however, one can see that this sort of defence is difficult to sustain. A claim that such dramatic systematic deception is unlikely suggests that you possess some means of working out how probable it is that it is true. But judgements of the probability of systematic deception will be based on experience just as prone to deception as any other. It is true that the brain in a vat scenario may seem unlikely to you, but that is just because it does not accord with your normal course of experience, and that experience is as open to the evil scientist's manipulation as any other. So you not only can't rule out the possibility of being a brain in a vat, you also can't say that that possibility is unlikely either. If you cannot even say that the brain in the vat scenario is unlikely, then the conclusion seems to be that you do not have *any* good reason for not thinking that you are a brain in a vat.

Arguments like this are often used to motivate scepticism concerning our beliefs about the external world. If we do not have a good reason for not thinking that we are brains in vats, then it seems we do not have a good reason not thinking that (nearly all) our beliefs about the external world are false. Since Descartes' suggestion in his *First Meditations*, this kind of radical doubt has been called 'Cartesian scepticism'. Not surprisingly, conclusions of this sort are unpalatable to many philosophers, and Cartesian scepticism has stimulated an enormously rich set of responses. In this paper, we wish to focus on just one: Putnam's discussion in his book *Reason, Truth and History* (Putnam 1981).

Putnam sets out to provide a very unusual argument against Cartesian scepticism. He argues that the compelling nature of the sceptical scenario derives from a misconceived view of the semantics of natural language. Once one appreciates the constraints on connections between language and the world, Putnam argues, one can appreciate that the sceptic's scenario is in a sense unstateable in natural language. Putnam's argument depends on demonstrating that the sentence 'I am a brain in a

vat' has a peculiar semantic property: when someone utters it, it is false *whether or not* that person is a brain in a vat.

This sounds odd, and bears repeating. Putnam, like everyone else, thinks that 'I am a brain in a vat' is false if I am not a brain in a vat. But surprisingly, Putnam thinks that 'I am a brain in a vat' is false (if it is meaningful at all) if I *am* a brain in a vat. Therefore, I cannot be a brain in a vat. Putnam's argument for this second, curious conclusion depends on an examination of the way that words interact with the world.

2 Reference and magic

The relationship of words to the world is a puzzling one. What makes a word about something? Consider the following ink-marks: 'Winston Churchill'. These ink-marks are not just any old set of marks. For many language users, they succeed in referring to a particular individual, namely, Winston Churchill. The marks are not at all like Churchill, who was a cigar-smoking politician. The marks do not smoke cigars. How is it possible that the set of marks that are black and squiggly come to be about something that gave speeches and smoked cigars? How does a set of physical marks get to *be about* something? This is one of the most fundamental questions in philosophy of language.

A tempting answer is that words are associated with our thoughts, in the form of mental representations, and it is mental representations that refer. This raises the question of how mental representations can refer, and it is not clear that this is any easier to answer than the original question with words. However, we will see that many constraints that apply to how words refer also apply to mental representations.

The relationship of *being about* that obtains between words and what they represent is called 'reference'. The pattern of ink-marks 'Winston Churchill' *refer* to Winston Churchill, *reference* to Winston Churchill is secured by the ink-marks. There are many theories of how reference works, and none of them is satisfactory. However, Putnam's argument is not focused on any particular theory of reference. He focuses on what is required at a minimum for reference to succeed—what needs to be satisfied according to any reasonable account of how reference works. Putnam's argument against Cartesian scepticism relies on discovering what these basic requirements are.

Putnam demonstrates these basic requirements with thought experiments that examine the success or failure of reference under different conditions. His strategy is to compare cases where an identical set of marks, or representations, succeed in referring in one case but fail in another. He argues that insight into the basic

requirements on reference is gained by looking at the difference between these two cases.

Putnam begins by considering an ant crawling across a patch of sand. It crawls around, going about its antlike business, and in the process, it traces out a likeness of Winston Churchill in the sand. Do the ant's tracks succeed in referring to Winston Churchill?

The ant has never seen Churchill, never seen a picture of Churchill, it had no intention of referring to Churchill, and it would have traced the same path if Churchill had never existed. It seems that it is simply an accident that the ant's tracks look like Churchill to those people who have seen pictures of him. It seems hard to say that the ant could even *be capable* of referring to Churchill. The ant just doesn't seem to have the right cognitive abilities to be capable of referring to Churchill. But what of the fact that the path still looks like Churchill? Doesn't the track the ant made succeed in referring to Churchill in virtue of its intrinsic Churchill-esque shape? This doesn't seem right either. Instead, it seems that we, as interpreting agents—agents who know who Churchill is and what he looked like—see the path as referring to Churchill. There is nothing *intrinsic* to the marks the ant made in the sand that make them refer to Churchill. In other words, the qualitative similarity of the path to what we regard as a picture of Churchill is not in itself a sufficient condition for reference.

The reason Putnam gives this example is to try to undermine the idea that there are ways of representing such that the mere presence of a symbol—the physical marks in the sand—by itself produces reference to an object. The idea is that some representational symbols have an *intrinsic* power to represent. Putnam calls theories of reference based on this idea 'magical' theories of reference.

It is not difficult to appreciate that there is nothing about words that makes them intrinsically about the things they represent. Apples are round, grow on trees, and are edible. The word 'apple', on the other hand, is none of these things. So it seems that the marks 'apple' do not intrinsically refer to apples (consider what would happen if someone who only spoke Swahili came across the page). Of course, this simple observation is not enough to show that magical theories of reference are wrong; Putnam's task is much more difficult than that. For example, one could argue that there are more fundamental candidates for representation. The most tempting candidates are *mental representations*. While it may be that there is nothing necessary about a set of marks in the sand referring to Winston Churchill, surely there is something necessary about a mental representation of Winston Churchill referring to Winston Churchill. Surely it is impossible that one could entertain a 'Winston Churchill' mental representation and yet still fail to refer to Churchill.

One could develop the suggestion that mental representations intrinsically represent

in a number of different ways. For example, one might think that each mental representation has a unique pattern of brain activity associated with it. There would be one distinctive pattern of brain activity associated with your mental representation of Winston Churchill, another distinctive pattern with your mental representation of Neville Chamberlain, a distinctive pattern again associated with your mental representation of Clement Atley, and so on. If mental representations do intrinsically refer, then the mere presence of these distinctive patterns of brain activity will be enough to guarantee reference to the respective British prime minister. For example, assume that a distinctive pattern of brain activity *W*, is sufficient for or constitutive of, your mental representation of Winston Churchill. Then *the mere physical presence* of the pattern *W* in your brain will be enough, by itself, to guarantee that your thought is about Winston Churchill.

Putnam claims that even if my Churchill mental representation is associated with a distinctive pattern of brain activity, the mere physical presence of that pattern is still not enough to guarantee reference to Churchill. In Putnam's view, it is possible both to entertain *all the qualitative aspects* of a mental representation, and to be in *the same physical state as one that refers*, and yet fail to refer. In other words, there is no magical link that guarantees that the mere presence of a distinctive pattern of brain activity, or distinctive mental experience, will produce reference. In order to support this claim, Putnam needs to make his argument on several different fronts: he needs to show that no matter what one thinks mental representations are—and there are very diverse views on this topic—those things cannot intrinsically represent. Putnam's most famous arguments against intrinsic mental representation appear in his discussion of the twin-Earth and elm and beech examples (Putnam 1975). However, it is perhaps easiest to appreciate his argument in the case of mental images, so we will focus on just that. For the purpose of the rest of the discussion, take it that mental representations are mental images, and take it on trust that Putnam's argument can be extended to other types of mental representation.

In order to defend his position, Putnam needs to show how it is possible for someone to have the exact same mental representation (mental image) that I do when I represent, for example, Winston Churchill, and yet they fail to refer to Churchill while I succeed. How is it possible to have an experience that is qualitatively identical to having a mental representation of Churchill and yet that experience fail to be about Churchill?

Putnam introduces another thought experiment. Suppose that there is a planet on which human beings have evolved, but suppose that these humans, although otherwise like us, have never seen or imagined trees. One day a picture of a tree is dropped on their planet by a passing spaceship. Imagine the inhabitants examining the picture. What could it possibly be a picture of? All sorts of speculations occur

to them: a building, a canopy, even an animal of some kind. Suppose that they never come close to the truth. *For us* the picture is a representation of a tree. For them the picture represents a strange object, nature and function unknown. Now, suppose that as a result of seeing the picture, one of these humans develops a mental image that is exactly the same as the mental image we have of trees. But, in the case of this other-worlder, the mental image is not a *representation of a tree*. It is a representation of a strange object, nature and function unknown.

One might argue that her mental image *is* a representation of a tree because there is a causal chain connecting her mental image back to the trees that the picture was made to represent. But this could be imagined absent. Suppose that the picture dropped from the spaceship is not really a picture of a tree but the accidental result of spilled paints. Even if it looked like a picture of a tree it would no more be a picture of a tree than the ant's picture of Churchill was a picture of Churchill. We can even assume that the spaceship that dropped the picture came from a planet that knew nothing of trees. It now seems hard to claim that the other-worlder's mental image succeeds in referring to trees, no matter how similar it is to your mental image. Mental images, just like physical pictures, do not have an intrinsic link to what they represent.

Let's assume that Putnam's argument can be generalised to other forms of mental representation, not just mental images. If this is the case, then two qualitatively identical individuals can have the same mental images, the same mental representations, even the same patterns of brain activity, yet one succeeds in thinking about trees while the other does not. The point is that nothing about any kind of representation—pictures, words, or mental images—intrinsically guarantees reference to something in the world. That can only be settled by things outside the representation. What could these external factors be? One has already been mentioned: causal contact.

We saw how a tree image resulting from an accidental spill of paints, in the hands of a community who has never seen or heard of trees, cannot refer to trees. This suggests something that might be a necessary condition for reference: there has to be a causal chain linking the representation to what it represents. If the chain is absent—as in the community who had never seen a tree—then it is hard to see how that community could ever refer to trees, no matter how much their mental representations resemble ours. This causal link needs to be of an appropriate kind, not just any causal link will do. For example, the fact that trees on distant planets gravitationally influence our tree-less community is not enough to enable them to refer to trees because it is not a sufficiently distinctive causal link to pick out trees as opposed to anything else. Putnam does not say a great deal about the nature of the appropriate causal link of reference, but he is clear that the causal link he has in mind is the one in virtue of which my thought 'tree' comes to be about actual

trees. The simplest way of understanding this is to imagine that my thought 'tree' is caused by me being direct perceptual contact with trees. A *tree* is what causes me to have a 'tree' thought, and it is that which guarantees that I refer to trees. Consider the other-world humans. Their 'tree-like' thought is caused by a random spill of paints, so there is nothing in virtue of which their thoughts could be about trees.

At this stage in Putnam's argument, he has concluded that reference depends on an appropriate kind of causal connection between a representation and what it represents. We now begin to get an idea of how he will take apart the brain in a vat hypothesis.

3 Brains in vats

A brain in a vat will have qualitatively the same experiences as a normal human being. It will have the same mental images, the same mental representations, and it will be in the same physical states (at least as far as its brain is concerned, since it has no body). However, it lacks one important thing: causal contact with the objects that it thinks exist in the world. Unlike a normal human, the brain in a vat is not in causal contact with any of the objects that we usually assume correspond to conscious experience: trees, flowers, books, tables, and so on. The brain in the vat is just in contact with its electrodes and its vat. Whatever else might exist in the real world need not be in causal contact with the brain. Even the evil scientist need not be there: the universe might consist of just automated machinery tending to a brain and stimulating it to have conscious experience.

The brain in a vat's mental life will look and feel the same as ours. Its mental images of trees will be qualitatively identical to our mental images of trees. Its experiences of trees will feel the same as our experiences of trees. It will even be able to use the word 'tree' in response to all same sorts of experiences as we do. But for all this similarity, Putnam argues, the brain in a vat cannot refer to trees. It has never been in causal contact with trees and so cannot mean *tree* by 'tree'. It can no more refer to trees than the inhabitants of the tree-less planet.

Let us turn from mental representations of trees to mental representations of brains and vats. Suppose that I, hopefully an embodied person, mentally represent a vat. Does my representation succeed in referring to vats? We can see that according to the way that Putnam has developed his account, it does. My mental representation of a vat is appropriately caused by actual interactions with vats. What about the brain in a vat? Does it refer to vats with its mental representation of 'vat'? The brain has never been in any kind of perceptual relationship with a real vat. Its representation of a vat is not related to real vats, even though its representation is qualitatively

just like mine. The brain's representation is caused by the stimuli that it receives from the computer that are vat-like, and it is in virtue of these that it comes to have vat-representations. At best, the brain refers to vat-computer-images when it has vat representations. But clearly when I say I am a brain in a vat, I do not mean that I am a brain-computer-image in a vat-computer-image, I mean that I am an actual brain in an actual vat, none of which I can successfully represent.

So, in spite of the fact that the brain in a vat can have qualitatively the same experiences as a normal human, it cannot refer to brains or vats. It cannot mean *brains* by 'brains', or *vat* by 'vat'. This is because it is not in appropriate causal relations with brains or vats—it has not, for example, ever been in perceptual contact with a brain or a vat. The brain in a vat is in a similar position to someone in our tree-less community. The members of that community failed to refer to trees, despite having tree-like mental images, because they were not in appropriate causal contact with trees. The brain in a vat, despite having the possibility of brain-like and vat-like experiences, cannot refer to real *brains* or *vats* because it is not in appropriate causal contact with them. If the brain in a vat cannot refer to real brains or real vats, then it cannot mean *I am a brain in a vat* by 'I am a brain in a vat'. Whatever it means when it says or thinks 'I am a brain in a vat', it cannot be *I am a brain in a vat*.

Perhaps one could argue that the situation concerning brains and vats is different to that with trees, since the brain in a vat *is* in causal contact with brains and vats—namely, with its own brain and its own vat. However, this sort of causal link does not seem sufficiently distinctive to enable reference to brains and vats as opposed to anything else. The same causal connection is present in all of the vat brain's conscious experiences. It is hard to see how it can do distinctive work in making the vat brain's representations of brains and vats refer to brains and vats.

Consider now the truth value of the statement 'I am a brain in a vat'. Let us suppose that the world is really as it appears, and I am an embodied person. In those circumstances, I am evidently not a brain in a vat, so 'I am a brain in a vat' is false. Suppose now that I am being dramatically deceived, and I am in fact a brain in a vat. What is the meaning of 'I am a brain in a vat' if not *I am a brain in a vat*? If the statement is not meaningless, then the most likely interpretation is that it refers to a state of affairs in the fictional world that the brain believes it inhabits. Now consider the truth value of the sentence 'I am a brain in a vat' under this interpretation. In the fictional world that the vat brain believes it inhabits it is not a brain in a vat—it is a living breathing human. So the sentence 'I am a brain in a vat' comes out false. Even if I am a brain in a vat, the statement 'I am a brain in a vat' comes out false.

This is the nub of Putnam's argument, the peculiar semantic property that we mentioned at the beginning: the sentence 'I am a brain in a vat' is false if uttered by someone who is not a brain in a vat, and false if uttered by someone who is a brain

in a vat. The sceptical hypothesis that we are brains in a vat, although it is perfectly consistent with everything that we have experienced and it violates no physical law, cannot possibly be true. I cannot be a brain in a vat.

This is bad news for the Cartesian sceptic. The Cartesian sceptic presented the brain in a vat hypothesis as something that might be true and then challenged us to refute it. Since the manipulations of the evil scientist were so subtle and systematic, it seemed that it would be impossible to answer the sceptic's challenge. But Putnam seems to have done just that. He has shown that no matter how the Cartesian sceptic phrases his hypothesis that hypothesis will be false. The sceptic's hypothesis is self-refuting.

Unfortunately, things are not so simple. Putnam's argument is ambitious and risks either proving too much or proving too little. It risks proving too much in that the causal contact condition seems hard to meet for many discourses. For example, it is seems hard to meet the condition for mathematical discourse or ethical discourse. The argument risks proving too little in that it seems possible modify the sceptic's hypothesis so as to escape the criticism while keeping its sceptical consequences. One could for example suggest an alternative sceptical hypothesis that you are a *post-operative* brain in a vat: you were once a living breathing human but now you are a brain in a vat. You can still refer to brains, vats, and everything else since your causal connection to those objects is secured via your causal links through your past (pre-operative) brain states. However, the sceptic could suggest that the real world and your simulated world have diverged so drastically since your operation that nearly all of your current beliefs are false. It seems that Cartesian scepticism is a coherent position after all.

4 Metaphysical realism and internal realism

Putnam's argument may not succeed in refuting Cartesian scepticism but, contrary to what we initially suggested, his argument is not primarily intended as a refutation of Cartesian scepticism anyway. Putnam's main target is not Cartesian scepticism but a view he calls 'metaphysical realism'. His brain in a vat argument is just one step on the road to discrediting metaphysical realism. Metaphysical realism is the view that there are mind-independent objects that we succeed in referring to with our language. Putnam discredits this view by arguing that no possible set of factors, either internal or external to us, could suffice in picking out mind-independent objects. There can be no relation of being about, or representation, between our words and mind-independent objects; there can be no reference relation between words and mind-independent objects. The brain in a vat argument is one step in this argument in that it tries to establish that no factor *internal* to us will do

the job of securing reference to mind-independent objects. No distinctive mental image, mental experience, mental representation, or brain state is enough to secure reference to those objects—the brain in a vat shares all these features with us yet, as Putnam argued, it fails to refer. At this point in his argument, Putnam still leaves it open that external factors such as causal connections could do the job of producing reference; however, he goes on to argue against this possibility. He claims that even if all possible external factors are taken into account there is still no way that a determinate reference-like relation to mind-independent objects can be secured. Our language and thought (or any imaginable variant thereof) simply does not have the semantic wherewithal to refer to mind-independent objects. Hence, in an important sense, *there are no mind-independent objects*—at least none that it makes sense to think or talk about. Metaphysical realism is denied.

On Putnam's view, we as embodied humans are in a very similar position to brains in vats. When we say 'tree' we do not somehow magically succeed in referring to mind-independent objects, *trees*. The brain in a vat could also not refer to *trees*. The most that the brain in the vat could do was "refer" to some part of the fictional world that it lived in: by 'tree' it meant *tree-computer-image*, or *tree-construct-on-the-computer*. Putnam's suggestion is that we are in the same the position. When we say 'tree' we do not succeed in referring to mind-independent objects *trees*, but "refer" to some (mind-dependent) part of our phenomenal world: we "refer" to *tree-phenomenon*, or *tree-appearance*. The things that we "refer" to are mind-dependent, they are other parts of our conceptual scheme. So, in this sense, none of the objects that we think and talk about are mind-independent. This goes for the objects of science—the earth, the sun, molecules, atoms, and quarks—as well as for trees and vats. Putnam calls this view 'internal realism'. For the internal realist, the mind-independent world is not pre-divided into facts and objects that we try to latch onto with our words. Rather, facts and objects as we know them are internal features of our conceptual scheme. Little or nothing can be said about the world beyond our conceptual scheme since no determinate reference relation to it can exist.

The contrast between metaphysical realism and internal realism is sometimes presented in terms of acceptance or rejection of a correspondence theory of truth. This is not entirely accurate because there are other ways of rejecting the correspondence theory of truth, but it can be helpful in distinguishing the positions. Metaphysical realism involves a commitment to the correspondence theory of truth. The correspondence theory of truth claims that a belief is true if and only if it stands in some correspondence relation to a mind-independent fact. Metaphysical realism holds that such mind-independent objects exist and the relation connecting them to beliefs is the reference relation. Putnam claims that no such reference relation can exist. He suggests that we should switch to internal realism. Internal realism denies both the existence of a correspondence relation and the existence of mind-

independent objects as we know them. This means that the internal realist needs to tell a different story about truth. Putnam suggests that truth is an idealised, aimed for, coherence in our conceptual scheme system: an ideal fit between our theories of the world and the facts as we experience and conceptualise them.

As it stands, Putnam's brain in the vat argument is compatible with metaphysical realism, indeed it presupposes a metaphysical realist perspective (it asks how the brain in vat can refer to mind-independent objects). But Putnam's strategy is to assume this perspective only in order to show that it fails by its own standards: he argues from within the metaphysical realist perspective that the reference-like relation that that perspective requires cannot exist. However, it is worth noting that an internal realist version of the brain in the vat argument can also be given. Both the metaphysical and internal realist reach the same conclusion that the sentence 'I am a brain in a vat' is necessarily false. We will conclude with a brief summary of an internalist version of the brain in a vat argument.

The reason why the brain in the vat hypothesis comes out false on the internal realist perspective is that the brains in the vat will conceptualise the world in a radically different from the way that we would if we were watching them from the outside. Suppose that the members of the current human population are nothing more than brains in a vat. What counts as facts and objects for these brains in the vat will be different from what we, as external observers, would count as facts and objects. Facts and objects for the brains in the vat are features of their fictional world; the facts that the brains in the vat experience and try to organise are facts in the fiction. This means that the sentence 'I am a brain in a vat' will, as Putnam originally argued, always come out false for them. This is because in the fiction the fact is that they are not brains in vats, but living breathing humans. Now, since in Putnam's thought experiment there is no one else to conceptualise the facts in any other way—and in particular no view from outside the vat—this means that there is no conceptual scheme under which the sentence 'I am a brain in a vat' will come out as true. So that sentence must be false. So the sentence 'I am a brain in a vat' is false whether one is a brain in a vat or not. Of course, this argument will have no force unless you already accept that facts and objects are mind-dependent in the way Putnam thinks they are. That is why he had to give a different argument in his discussion of the brain in the vat. However, it is reassuring that both perspectives reach the same conclusion. For Putnam, both the metaphysical and the internal realist can agree that we cannot possibly be brains in a vat.

References

Putnam, H. 1975. 'The Meaning of "Meaning"'. In *Mind, Language and Reality, Philosophical Papers, vol. 2,* 215–271. Cambridge: Cambridge University Press.

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