48

Is There a Possibility of Being a Mere and Unknowing Brain in a Vat?

Hilary Putnam, *Reason, Truth and History*

David Lewis was not arguing (in the previous reading) directly against skeptical thinking. But perhaps an aspect of his approach may be used for that end. Lewis sought to understand what *connection* to the world must be part of really seeing the world. This reading from Hilary Putnam (b. 1926) argues from a related picture, of a connection to the world that would be part even of *apparently* seeing an aspect of the world. Putnam's career, mainly at Harvard University, has been very influential. He has formulated many theories and thought-experiments about language, logic, reality, mind, and more. This much-discussed argument of his is anti-skeptical.

It is directed against a contemporary version – a re-imaging – of Descartes' dreaming argument. This newer version is usually called the *brain-in-a-vat* skeptical argument. Like the dreaming argument, this one raises a possibility in which someone has a pattern of experiences that feel, "from within," like normal perceptual interactions with the physical world – even though in fact they are far from normal. But what makes this skeptical argument distinctive is a further detail. This time, the conscious experiences belong to a *disembodied brain*, floating in a vat of sustaining chemicals. The brain is attached electrically to a machine causing it to have those experiences: the "inner" experiences had by the brain in the vat are programmed through the machine. That is the possibility envisaged by this argument. Then the skeptical argument raises this question: Can you know that you are *not* in that predicament? Indeed, can you know that you have not been in it for a long time? Can you know that you are not merely a long-term brain in a vat – hence, that for a long time you have not actually been observing the world, even when feeling like you are observing it?

Hilary Putnam, *Reason, Truth and History* (Cambridge: Cambridge University Press, 1981), ch. 1 (excerpts).

Metaphysics and Epistemology: A Guided Anthology, First Edition. Edited by Stephen Hetherington.

Putnam's reply asks whether the supposed skeptical possibility really is possible. Any experience is the experience it is partly because of its *content*. Yet how would those experiences belonging to the hypothesized long-term brain in the vat *have* a particular content? Putnam asks generally about how an apparently subjectively "owned" experience gains its content (e.g. "That's a fox"). For him, this is a question about how a word such as "fox" refers or denotes, how it is *about* something at all. What is needed for real reference (argues Putnam) is an apt *history of causal interaction* between uses of the word and elements of the world: the word "fox" has its meaning partly by having been used in responses to foxes. But no long-term brain in a vat could have been having that interaction. The long-term brain in the vat has not been responding to foxes. It has merely been "fed" apparently sensory experiences by the machine.

Significantly (continues Putnam), in such a circumstance you could not have had the sort of interaction required for your uses even of the phrase "brain in a vat" to contribute meaningfully to your thoughts. If you were a long-term brain in a vat, you would have been so systematically and sustainedly deceived about your surroundings that your uses of words would not mean what they seem to you to mean. You could not even think to yourself – with real content – "I *might* have long been a brain in a vat." So you cannot really entertain the possibility of being a long-term brain in a vat. Putnam thus claims to reveal as *self-refuting* any attempt to posit the skeptical possibility. Much is at stake in whether he is correct about this.

Brains in a Vat

An ant is crawling on a patch of sand. As it crawls, it traces a line in the sand. By pure chance the line that it traces curves and recrosses itself in such a way that it ends up looking like a recognizable caricature of Winston Churchill. Has the ant traced a picture of Winston Churchill, a picture that *depicts* Churchill?

Most people would say, on a little reflection, that it has not. The ant, after all, has never seen Churchill, or even a picture of Churchill, and it had no intention of depicting Churchill. It simply traced a line (and even *that* was unintentional), a line that *we* can 'see as' a picture of Churchill.

We can express this by saying that the line is not 'in itself' a representation[1] of anything rather than anything else. Similarity (of a certain very complicated sort) to the features of Winston Churchill is not sufficient to make something represent or refer to Churchill. Nor is it necessary: in our community the printed shape 'Winston Churchill', the spoken words 'Winston Churchill', and many other things are used to represent Churchill (though not pictorially), while not having the sort of similarity to Churchill that a picture – even a line drawing – has. If *similarity* is not necessary or sufficient to make something represent something else, how can *anything* be necessary or sufficient for this purpose? How on earth can one thing represent (or 'stand for', etc.) a different thing?

The answer may seem easy. Suppose the ant had seen Winston Churchill, and suppose that it had the intelligence and skill to draw a picture of him. Suppose it produced the caricature *intentionally*. Then the line would have represented Churchill.

On the other hand, suppose the line had the shape WINSTON CHURCHILL. And suppose this was just accident (ignoring the improbability involved). Then the 'printed shape' WINSTON CHURCHILL would *not* have represented Churchill, although that printed shape does represent Churchill when it occurs in almost any book today.

So it may seem that what is necessary for representation, or what is mainly necessary for representation, is *intention*.

But to have the intention that *anything*, even private language (even the words 'Winston Churchill' spoken in my mind and not out loud), should *represent* Churchill, I must have been able to *think about* Churchill in the first place. If lines in the sand, noises, etc., cannot 'in themselves' represent anything, then how is it that thought forms can 'in themselves' represent anything? Or can they? How can thought reach out and 'grasp' what is external?

[…]

Magical Theories of Reference

We saw that the ant's 'picture' has no necessary connection with Winston Churchill. The mere fact that the 'picture' bears a 'resemblance' to Churchill does not make it into a real picture, nor does it make it a representation of Churchill. […]

What is important to realize is that what goes for physical pictures also goes for mental images, and for mental representations in general; mental representations no more have a necessary connection with what they represent than physical representations do. The contrary supposition is a survival of magical thinking.

Perhaps the point is easiest to grasp in the case of mental *images*. (Perhaps the first philosopher to grasp the enormous significance of this point, even if he was not the first to actually make it, was Wittgenstein.) Suppose there is a planet somewhere on which human beings have evolved (or been deposited by alien spacemen, or what have you). Suppose these humans, although otherwise like us, have never seen *trees*. Suppose they have never imagined trees (perhaps vegetable life exists on their planet only in the form of molds). Suppose one day a picture of a tree is accidentally dropped on their planet by a spaceship which passes on without having other contact with them. Imagine them puzzling over the picture. What in the world is this? All sorts of speculations occur to them: a building, a canopy, even an animal of some kind. But suppose they never come close to the truth.

For *us* the picture is a representation of a tree. For these humans the picture only represents a strange object, nature and function unknown. Suppose one of them has a mental image which is exactly like one of my mental images of a tree as a result of having seen the picture. His mental image is not a *representation of a tree*. It is only a representation of the strange object (whatever it is) that the mysterious picture represents.

Still, someone might argue that the mental image is *in fact* a representation of a tree, if only because the picture which caused this mental image was itself a representation of a tree to begin with. There is a causal chain from actual trees to the mental image even if it is a very strange one.

But even this causal chain can be imagined absent. Suppose the 'picture of the tree' that the spaceship dropped was not really a picture of a tree, but the accidental result of some spilled paints. Even if it looked exactly like a picture of a tree, it was, in truth, no more a picture of a tree than the ant's 'caricature' of Churchill was a picture of Churchill. We can even imagine that the spaceship which dropped the 'picture' came from a planet which knew nothing of trees. Then the humans would still have mental images qualitatively identical with my image of a tree, but they would not be images which represented a tree any more than anything else.

The same thing is true of *words*. A discourse on paper might seem to be a perfect description of trees, but if it was produced by monkeys randomly hitting keys on a typewriter for millions of years, then the words do not refer to anything. If there were a person who memorized those words and said them in his mind without understanding them, then they would not refer to anything when thought in the mind, either.

Imagine the person who is saying those words in his mind has been hypnotized. Suppose the words are in Japanese, and the person has been told that he understands Japanese. Suppose that as he thinks those words he has a 'feeling of understanding'. (Although if someone broke into his train of thought and asked him what the words he was thinking *meant*, he would discover he couldn't say.) Perhaps the illusion would be so perfect that the person could even fool a Japanese telepath! But if he couldn't use the words in the right contexts, answer questions about what he 'thought', etc., then he didn't understand them.

By combining these science fiction stories I have been telling, we can contrive a case in which someone thinks words which are in fact a description of trees in some language *and* simultaneously has appropriate mental images, but *neither* understands the words *nor* knows what a tree is. We can even imagine that the mental images were caused by paint-spills (although the person has been hypnotized to think that they are images of something appropriate to his thought – only, if he were asked, he wouldn't be able to say of what). And we can imagine that the language the person is thinking in is one neither the hypnotist nor the person hypnotized has ever heard of – perhaps it is just coincidence that these 'nonsense sentences', as the hypnotist supposes them to be, are a description of trees in Japanese. In short, everything passing before the person's mind might be qualitatively identical with what was passing through the mind of a Japanese speaker who was *really* thinking about trees – but none of it would refer to trees.

[…] [E]ven a large and complex system of representations, both verbal and visual, still does not have an *intrinsic*, built-in, magical connection with what it represents – a connection independent of how it was caused and what the dispositions of the speaker or thinker are. And this is true whether the system of representations (words and images, in the case of the example) is physically realized – the words are written or spoken, and the pictures are physical pictures – or only realized in the mind. Thought words and mental pictures do not *intrinsically* represent what they are about.

The Case of the Brains in a Vat

Here is a science fiction possibility discussed by philosophers: imagine that a human being (you can imagine this to be yourself) has been subjected to an operation by an evil scientist. The person's brain (your brain) has been removed from the body and placed in a vat of nutrients which

keeps the brain alive. The nerve endings have been connected to a super-scientific computer which causes the person whose brain it is to have the illusion that everything is perfectly normal. There seem to be people, objects, the sky, etc; but really all the person (you) is experiencing is the result of electronic impulses travelling from the computer to the nerve endings. The computer is so clever that if the person tries to raise his hand, the feedback from the computer will cause him to 'see' and 'feel' the hand being raised. Moreover, by varying the program, the evil scientist can cause the victim to 'experience' (or hallucinate) any situation or environment the evil scientist wishes. He can also obliterate the memory of the brain operation, so that the victim will seem to himself to have always been in this environment. It can even seem to the victim that he is sitting and reading these very words about the amusing but quite absurd supposition that there is an evil scientist who removes people's brains from their bodies and places them in a vat of nutrients which keep the brains alive. The nerve endings are supposed to be connected to a super-scientific computer which causes the person whose brain it is to have the illusion that …

When this sort of possibility is mentioned in a lecture on the Theory of Knowledge, the purpose, of course, is to raise the classical problem of scepticism with respect to the external world in a modern way. (*How do you know you aren't in this predicament?*) But this predicament is also a useful device for raising issues about the mind/world relationship.

Instead of having just one brain in a vat, we could imagine that all human beings (perhaps all sentient beings) are brains in a vat (or nervous systems in a vat in case some beings with just a minimal nervous system already count as 'sentient'). Of course, the evil scientist would have to be outside – or would he? Perhaps there is no evil scientist, perhaps (though this is absurd) the universe just happens to consist of automatic machinery tending a vat full of brains and nervous systems.

This time let us suppose that the automatic machinery is programmed to give us all a *collective* hallucination, rather than a number of separate unrelated hallucinations. Thus, when I seem to myself to be talking to you, you seem to yourself to be hearing my words. Of course, it is not the case that my words actually reach your ears – for you don't have (real) ears, nor do I have a real mouth and tongue. Rather, when I produce my words, what happens is that the efferent impulses travel from my brain to the computer, which both causes me to 'hear' my own voice uttering those words and 'feel' my tongue moving, etc., and causes you to 'hear' my words, 'see' me speaking, etc. In this case, we are, in a sense, actually in communication. I am not mistaken about your real existence (only about the existence of your body and the 'external world', apart from brains). From a certain point of view, it doesn't even matter that 'the whole world' is a collective hallucination; for you do, after all, really hear my words when I speak to you, even if the mechanism isn't what we suppose it to be. (Of course, if we were two lovers making love, rather than just two people carrying on a conversation, then the suggestion that it was just two brains in a vat might be disturbing.)

I want now to ask a question which will seem very silly and obvious (at least to some people, including some very sophisticated philosophers), but which will take us to real philosophical depths rather quickly. Suppose this whole story were actually true. Could we, if we were brains in a vat in this way, *say* or *think* that we were?

I am going to argue that the answer is 'No, we couldn't.' In fact, I am going to argue that the supposition that we are actually brains in a vat, although it violates no physical law, and is perfectly consistent with everything we have experienced, cannot possibly be true. *It cannot possibly be true*, because it is, in a certain way, self-refuting.

[…]

A 'self-refuting supposition' is one whose truth implies its own falsity. For example, consider the thesis that *all general statements are false*. This is a general statement. So if it is true, then it must be false. Hence, it is false. Sometimes a thesis is called 'self-refuting' if it is *the supposition that the thesis is entertained or enunciated* that implies its falsity. For example, 'I do not exist' is self-refuting if thought by *me* (for any '*me*'). So one can be certain that one oneself exists, if one thinks about it (as Descartes argued).

What I shall show is that the supposition that we are brains in a vat has just this property. If we can consider whether it is true or false, then it is not true (I shall show). Hence it is not true.

Before I give the argument, let us consider why it seems so strange that such an argument can be given (at least to philosophers who subscribe to a 'copy' conception of truth). We conceded that it is compatible with physical law that there should be a world in which all sentient beings are brains in a vat. As philosophers say, there is a 'possible world' in which all sentient beings are brains in a vat. […] The humans in that possible world have exactly the same experiences that *we* do. They think the same thoughts we do (at least, the same words, images, thought-forms, etc., go through their minds). Yet, I am claiming that there is an argument we can give that shows we are not brains in a vat. How can there be? And why couldn't the people in the possible world who really *are* brains in a vat give it too?

The answer is going to be (basically) this: although the people in that possible world can think and 'say' any words we can think and say, they cannot (I claim) *refer* to what we can refer to. In particular, they cannot think or say that they are brains in a vat (*even by thinking 'we are brains in a vat'*).

[…]

[…] [T]here is the *illusion* that the ant has caricatured Churchill. […] [T]he ant would have drawn the same curve even if Winston Churchill had never existed. […]

Brains in a Vat (Again)

[…] The brains in a vat do not have sense organs, but they do have *provision* for sense organs; that is, there are afferent nerve endings, there are inputs from these afferent nerve endings, and these inputs figure in the 'program' of the brains in the vat just as they do in the program of our brains. The brains in a vat are *brains*; moreover, they are *functioning* brains, and they function by the same rules as brains do in the actual world. For these reasons, it would seem absurd to deny consciousness or intelligence to them. But the fact that they are conscious and intelligent does not mean that their words refer to what our words refer. The question we are interested in is this: do their verbalizations containing, say, the word 'tree' actually refer to *trees?* More generally: can they refer to *external* objects at all? […]

To fix our ideas, let us specify that the automatic machinery is supposed to have come into existence by some kind of cosmic chance or coincidence (or, perhaps, to have always existed). In this hypothetical world, the automatic machinery itself is supposed to have no intelligent creator–designers. […]

This assumption does not help. For there is no connection between the *word* 'tree' as used by these brains and actual trees. They would still use the word 'tree' just as they do, think just

the thoughts they do, have just the images they have, even if there were no actual trees. Their images, words, etc., are qualitatively identical with images, words, etc., which do represent trees in *our* world; but we have already seen (the ant again!) that qualitative similarity to something which represents an object (Winston Churchill or a tree) does not make a thing a representation all by itself. In short, the brains in a vat are not thinking about real trees when they think 'there is a tree in front of me' because there is nothing by virtue of which their thought 'tree' represents actual trees.

[…] [W]e have seen that the words do not necessarily refer to trees even if they are arranged in a sequence which is identical with a discourse which (were it to occur in one of our minds) would unquestionably *be about trees* in the actual world. Nor does the 'program', in the sense of the rules, practices, dispositions of the brains to verbal behavior, necessarily refer to trees or bring about reference to trees through the connections it establishes between words and words, or *linguistic* cues and *linguistic* responses. If these brains think about, refer to, represent trees (real trees, outside the vat), then it must be because of the way the 'program' connects the system of language to *non-verbal* input and outputs. There are indeed such non-verbal inputs and outputs in the Brain-in-a-Vat world (those efferent and afferent nerve endings again!), but we also saw that the 'sense-data' produced by the automatic machinery do not represent trees (or anything external) even when they resemble our tree-images exactly. Just as a splash of paint might resemble a tree picture without *being* a tree picture, so, we saw, a 'sense datum' might be qualitatively identical with an 'image of a tree' without being an image of a tree. How can the fact that, in the case of the brains in a vat, the language is connected by the program with sensory inputs which do not intrinsically or extrinsically represent trees (or anything external) possibly bring it about that the whole system of representations, the language-in-use, *does* refer to or represent trees or anything external?

The answer is that it cannot. The whole system of sense-data, motor signals to the efferent endings, and verbally or conceptually mediated thought connected by 'language entry rules' to the sense-data (or whatever) as inputs and by 'language exit rules' to the motor signals as outputs, has no more connection to *trees* than the ant's curve has to Winston Churchill. Once we see that the *qualitative similarity* (amounting, if you like, to qualitative identity) between the thoughts of the brains in a vat and the thoughts of someone in the actual world by no means implies sameness of reference, it is not hard to see that there is no basis at all for regarding the brain in a vat as referring to external things.

The Premisses of the Argument

I have now given the argument promised to show that the brains in a vat cannot think or say that they are brains in a vat. It remains only to make it explicit and to examine its structure.

By what was just said, when the brain in a vat (in the world where every sentient being is and always was a brain in a vat) thinks 'There is a tree in front of me', his thought does not refer to actual trees. On some theories that we shall discuss it might refer to trees in the image, or to the electronic impulses that cause tree experiences, or to the features of the program that are responsible for those electronic impulses. These theories are not ruled out by what was just said, for there is a close causal connection between the use of the word 'tree' in vat-English and the presence of trees in the image, the presence of electronic impulses of a certain

kind, and the presence of certain features in the machine's program. On these theories the brain is *right*, not *wrong* in thinking 'There is a tree in front of me.' Given what 'tree' refers to in vat-English and what 'in front of' refers to, assuming one of these theories is correct, then the truth-conditions for 'There is a tree in front of me' when it occurs in vat-English are simply that a 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. And these truth-conditions are certainly fulfilled.

By the same argument, 'vat' refers to vats in the image in vat-English, or something related (electronic impulses or program features), but certainly not to real vats, since the use of 'vat' in vat-English has no causal connection to real vats (apart from the connection that the brains in a vat wouldn't be able to use the word 'vat', if it were not for the presence of one particular vat – the vat they are in; but this connection obtains between the use of *every* word in vat-English and that one particular vat; it is not a special connection between the use of the *particular* word 'vat' and vats). Similarly, 'nutrient fluid' refers to a liquid in the image in vat-English, or something related (electronic impulses or program features). It follows that if their 'possible world' is really the actual one, and we are really the brains in a vat, then what we now mean by 'we are brains in a vat' is that *we are brains in a vat in the image* or something of that kind (if we mean anything at all). But part of the hypothesis that we are brains in a vat is that we aren't brains in a vat in the image (i.e. what we are 'hallucinating' isn't that we are brains in a vat). So, if we are brains in a vat, then the sentence 'We are brains in a vat' says something false (if it says anything). In short, if we are brains in a vat, then 'We are brains in a vat' is false. So it is (necessarily) false.

[…]

[…] Concepts are signs used in a certain way; the signs may be public or private, mental entities or physical entities, but even when the signs are 'mental' and 'private', the sign itself apart from its use is not the concept. And signs do not themselves intrinsically refer.

We can see this by performing a very simple thought experiment. Suppose you are like me and cannot tell an elm tree from a beech tree. We still say that the reference of 'elm' in my speech is the same as the reference of 'elm' in anyone else's, viz. elm trees, and that the set of all beech trees is the extension of 'beech' (i.e. the set of things the word 'beech' is truly predicated of) both in your speech and my speech. Is it really credible that the difference between what 'elm' refers to and what 'beech' refers to is brought about by a difference in our *concepts?* My concept of an elm tree is exactly the same as my concept of a beech tree (I blush to confess). (This shows that the determination of reference is social and not individual, by the way; you and I both defer to experts who *can* tell elms from beeches.) If someone heroically attempts to maintain that the difference between the reference of 'elm' and the reference of 'beech' in *my* speech is explained by a difference in my psychological state, then let him imagine a Twin Earth where the words are switched. Twin Earth is very much like Earth; in fact, apart from the fact that 'elm' and 'beech' are interchanged, the reader can suppose Twin Earth is exactly like Earth. Suppose I have a *Doppelganger* on Twin Earth who is molecule for molecule identical with me (in the sense in which two neckties can be 'identical'). If you are a dualist, then suppose my *Doppelganger* thinks the same verbalized thoughts I do, has the same sense data, the same dispositions, etc. It is absurd to think his psychological

state is one bit different from mine: yet his word 'elm' represents *beeches*, and my word 'elm' represents *elm*. (Similarly, if the 'water' on Twin Earth is a different liquid – say, XYZ and not H_2O – then 'water' represents a different liquid when used on Twin Earth and when used on Earth, etc.) Contrary to a doctrine that has been with us since the seventeenth century, *meanings just aren't in the head*.

Note

1. The terms 'representation' and 'reference' always refer to a relation between a word (or other sort of sign, symbol, or representation) and something that actually exists (i.e. not just an 'object of thought'). There is a sense of 'refer' in which I can 'refer' to what does not exist; this is not the sense in which 'refer' is used here. An older word for what I call 'representation' or 'reference' is *denotation*.