Phil. 159: Epistemology
Feb. 15, 2022

Lecture 7: Putnam on Brains-in-Vats

I. Putnam against Magical Theories of Reference

Suppose an ant traces a line in the sand that just happens to look exactly like a caricature of Winston Churchill. Has the ant traced a *picture* that *depicts* (or *represents*) Winston Churchill?

> *Putnam's answer*: No, it hasn't.

Suppose that same ant traces a line in the sand that just happens to take the shape 'WINSTON CHURCHILL'. Has the ant traced a *word* that *refers to* Winston Churchill?

> *Putnam's answer*: No, it hasn't.

Suppose a race of humans who have evolved on a distant, treeless planet discover a picture that looks exactly like a picture of a tree. Suppose, also, that the picture was the accidental result of some spilled paint. Finally, suppose that, as result of having seen that picture, one of these humans has a mental image that is exactly like one of our mental images of trees. Does she thereby have a *mental image* that *represents* a tree?

> *Putnam's answer*: No, she doesn't.

Suppose a person has been hypnotized to repeat in her head the Japanese words that mean "There is a tree before me." Suppose, also, that as she thinks these words, this person has a "feeling of understanding." Does she thereby have a *thought* that *refers to* a tree?

> *Putnam's answer*: No, she doesn't.

Putnam provides the following diagnosis for why reference fails in all of these sorts of cases:

> *Putnam's causal constraint on reference*: Pictures, words, mental images, and thoughts cannot refer to individuals (such as Winston Churchill) or natural kinds (such as trees) unless the appropriate sort of causal connection obtains between the picture/word/image/thought and the individual/kind in question.

Some difficult questions that we don't need to take a stand on here:

- What sort of causal connection is needed?
- Is the appropriate sort of causal connection *sufficient* for reference (in addition to being *necessary*)?
- What other kinds of terms besides *names* and *natural-kind terms* are subject to such a causal constraint?

Theories of reference that violate the causal constraint (and hence hold that words/images/thoughts have an intrinsic, built-in connection with what they represent) Putnam disparagingly calls *magical theories of reference*.

II. Variants of the Brain-in-a-Vat Scenario

The standard brain-in-a-vat scenario can vary along several dimensions:

> Is there only one brain-in-a-vat (*solipsistic version*), or are there many brains-in-vats arranged so as to have a collective hallucination (*collective version*)?
>
> Have the brains-in-vats been recently envatted (*recent version*), or have they been envatted for their entire lives (*eternal version*)?
>
> Is there an evil scientist who created the vat technology (*evil-scientist version*), or does the universe just happen to consist of automatic machinery tending various brains in vats (*automatic-machinery version*)?

Putnam will be considering the *collective, eternal, automatic-machinery version* of the brain-in-a-vat scenario. (Henceforth, let 'brain-in-a-vat' refer to this sort of a brain-in-a-vat.)

III. Putnam's Argument

According to Putnam, if we were brains-in-vats, then we couldn't say or think that we were brains-in-vats, and thus the supposition that we are brains-in-vats "*cannot possibly be true*" (p. 7).

Putnam insists that when a brain-in-a-vat thinks, "There is a tree in front of me," her thought doesn't refer to actual trees because the right sort of causal connection doesn't obtain between the brain-in-a-vat and trees.

> What *does* her thought refer to? Putnam is non-committal: "On some theories . . . 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 electronic impulses" (p. 14). Later he even countenances the possibility that her thought might not refer to anything at all.
>
> However, Putnam's argument is easier to gloss if we assume a certain account of what the brain-in-a-vat's thoughts refer to. So for the time being let us assume that 'tree' in the mouth of a brain-in-a-vat refers to a "tree-in-the-image" (i.e. some stable conglomerate of tree-like sense impressions).

Working under this assumption, Putnam insists that when a brain-in-a-vat says or thinks the word 'brain', she refers to a brain-in-the-image, not a real brain. Similarly for 'vat' and 'nutrient fluid'.

But then it appears we can argue as follows (pp. 14-15):

Putnam's argument (on one interpretation):

1. If I am a brain-in-a-vat, then the sentence "I am a brain-in-a-vat" (as thought or said by me) is true if and only if I am a brain-in-a-vat-in-the-image. [*premise*]
2. If I am a brain-in-a-vat, then I am not a brain-in-a-vat-in-the-image. [*premise*]
3. So, if I am a brain-in-a-vat, then the sentence "I am a brain-in-a-vat" (as thought or said by me) is false. [*follows from 1, 2*]
4. If I am not a brain-in-a-vat, then the sentence "I am a brain-in-a-vat" (as thought or said by me) is false. [*premise*]
5. So, the sentence "I am a brain-in-a-vat" (as thought or said by me) is false. [*follows from 3, 4*]
6. So, I am not a brain-in-a-vat. [*follows from 5*]

This formulation of the argument assumes a theory of reference according to which 'brain-in-a-vat' in Vat-English refers to a brain-in-a-vat-in-the-image. However, we can discharge that assumption: Putnam thinks that any theory of reference which obeys his causal constraint will yield the result that (3) holds.

> *worry #1a*: Even if Putnam is right that a brain-in-a-vat can't talk or think about her predicament, so that there's a certain sense in which *we* couldn't talk or think about our predicament if we *were* brains-in-vats, doesn't it still seem possible that we *are* in that predicament (or one sufficiently similar to it), but unable to refer to our lot by saying, "We are brains-in-vats"?
>
> *worry #1b*: Indeed, it appears we might be able to refer to our predicament after all, if we assiduously avoid using terms that satisfy Putnam's causal constraint. For example, maybe the possibility we are trying to entertain can be glossed as follows: "I am an isolated cognitive system that receives its inputs from and sends its outputs to an accidentally created computer simulation of a world."
>
> *worry #2*: Putnam's argument only seems to apply to the possibility that we are *eternally envatted* brains-in-vats. If we were envatted 10 minutes ago, then presumably our terms 'tree', 'brain', and 'vat' would refer to real trees, brains, and vats.
>
> > (Note, though, that *the recent brain-in-a-vat scenario* is less simple an hypothesis than *the real-world scenario* or *the eternal brain-in-a-vat scenario*. So maybe Putnam's argument can be combined with an inference-to-the-best-explanation response to Descartes' argument to refute skepticism.)