Multiple realizability and functionalism

PHIL 30304
Jeff Speaks

September 4, 2018

1 The argument from multiple realizability

Putnam begins 'The nature of mental states' by agreeing with a lot of claims that we saw Smart making. Putnam agrees with Smart that it is coherent to think of the identification of pains and other mental states with brain states as the same kind of claim as other theoretical identifications in science, and agrees further that the fact that one can know that one is in pain without knowing much about one's brain state does not show that pains $\neq$ brain states. (As Putnam points out, and as Smart did, if this argument was good it would count against almost any scientific theoretical identification.)

His argument against the view is not that it is nonsense, but that when we look at what it would take for the identity theory to be true, we can see that it is very unlikely to be true:

> "Consider what the brain state theorist has to do to make good his claims. He has to specify a physical-chemical state such that *any* organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain ... etc. At the same time it must *not* be a possible ... state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extraterrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain.
>
> ...
>
> Finally, the hypothesis becomes still more ambitious when we realize that the brain-state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus ... but whose physical-chemical 'correlate' is different in the two cases, the brain-state theory has collapsed." (436-7)

Arguments of this form against identity theories of mental states are sometimes called arguments from 'multiple realizability.' The core assumption is the assumption that very

different kinds of creatures might have certain mental properties in common — like feeling a certain pain — but not have the relevant physical properties in common.

One might state Putnam's argument from multiple realizability against the identity theory as follows, letting M be some arbitrary mental property:

1. Distinct creatures both have M despite having no interesting physical-chemical properties in common.
2. If distinct creatures both have M despite having no interesting physical-chemical properties in common, then M is not identical to any physical-chemical property.

C. M is not identical to any physical-chemical property. (1,2)

How should the identity theorist respond? How would Smart reply?

A different version of the argument from multiple realizability starts not with the assumption that different actual organisms have the same mental state without having the relevant physical-chemical properties in common, but with the weaker assumption that it is *possible* that distinct creatures have the same mental state without having the relevant physical-chemical properties in common. Surely, after all, it is possible that some alien made of silicone feel pain. That version of the argument might be stated like this:

1. It is possible that distinct creatures both have M despite having no interesting physical-chemical properties in common.
2. If it is possible that distinct creatures both have M despite having no interesting physical-chemical properties in common, then M is not identical to any physical-chemical property.

C. M is not identical to any physical-chemical property. (1,2)

How would Smart respond to this version of the argument?

2 Functionalism: a new theory of mental properties

Though he rejects the identity theory, Putnam does not think of mental properties as non-physical properties of an immaterial soul (like Descartes) or as logical constructions out of behavior (like Ryle). Instead, he introduces a new theory of mental states, which he expresses as the view that a mental state like pain is 'a functional state of a whole organism.' (433)

What does he mean by this? He says:

> 'To explain this it is necessary to introduce some technical notions . . . a Probabilistic Automaton is defined [as a machine such that] the transitions between states are allowed to be with various probabilities rather than being 'deterministic.' . . . I shall assume the notion of a Probabilistic Automaton has been generalized to allow for 'sensory inputs' and 'motor outputs' — that is, the Machine Table specifies, for every possible combination of a 'state' and a complete set of 'sensory inputs' an 'instruction' which determines the probability of the next 'state,' and also the probabilities of the 'motor outputs.' . . . I shall

> also assume that the physical realization of the sense organs responsible for the various inputs, and of the motor organs, is specified, but that the 'states' and the 'inputs' themselves are, as usual, specified only 'implicitly' — i.e. by the set of transition probabilities given by the Machine Table. '

Are you a probabilistic automaton, in this sense?

Putnam's idea is that whether you have a certain mental property is determined by what sort of probabilistic automaton you are. To see how, exactly, this is supposed to work, it will be useful to have a look at the way functionalism is introduced in David Lewis' "Psychophysical and theoretical identifications" (on the web site as an optional reading). Lewis introduces functionalism via his example of the detective story:

> "We are assembled in the drawing room of the country house; the detective reconstructs the crime. That is, he proposes a theory designed to be the best explanation of phenomena we have observed: the death of Mr. Body, the blood on the wallpaper, the silence of the dog in the night, the clock seventeen minutes fast, and so on. He launches into his story:
>
> > X, Y and Z conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, X was Body's partner... Last week, Y and Z conferred in a bar in Reading... Tuesday night at l1:17, Y went to the attic and set a time bomb... Seventeen minutes later, X met Z in the billiard room and gave him the lead pipe... Just when the bomb went off in the attic, X fired three shots into the study through the French windows...
>
> And so it goes: a long story. Let us pretend that it is a single long conjunctive sentence.
>
> The story contains the three names 'X', 'Y' and 'Z'. The detective uses these new terms without explanation, as though we knew what they meant. But we do not. We never used them before, at least not in the senses they bear in the present context. All we know about their meanings is what we gradually gather from the story itself." (250)

The point of this is that there is a sense in which the story describes, or purports to describe, three people. 'X' stands for whoever did the stuff the story ascribes to 'X', 'Y' stands for whoever did the stuff the story ascribes to 'Y' etc. Another way to put that is that there is a certain *role* in the story corresponding to each of these letters. For the letter to stand for a person is for the person to *realize* that role:

> "Suppose that after we have heard the detective's story, we learn that it is true of a certain three people: Plum, Peacock and Mustard. If we put the name 'Plum' in place of 'X', 'Peacock' in place of 'Y', and 'Mustard' in place of 'Z' throughout, we get a true story about the doings of those three people. We will say that Plum, Peacock and Mustard together realize (or are a realization of) the detective's theory." (251)

Lewis's idea is that words for mental states, like 'feels pain' and 'believes that there is beer in the fridge', are like the letters in the detective story: they stand for *whatever state realizes a certain role*.

In the case of mental properties, in place of a detective story we have a story about the connections between sensory inputs, mental states, and behavioral outputs. Suppose, for example, we are interested in the mental property *believes that there is beer in the fridge*. Then our story might include claims like the following:

> If someone is placed in front of an open refrigerator which has beer in it, then he will believe that there is beer in the refrigerator.
>
> If someone wants beer and believes that there is beer in the refrigerator, then he will go to the refrigerator and get a beer.
>
> If someone believes that there is beer in the refrigerator, then he believes that there is beer somewhere.
>
> If someone believes that there is a Budweiser in the refrigerator, then he believes that there is beer in the refrigerator.
>
> If someone intends to get a beer out of the refrigerator, then he believes that there is a beer in the refrigerator.

If we think of claims like these as comprising a (somewhat boring) story, then the story has a number of 'characters'. One of these characters is the belief that there is beer in the refrigerator:

> If someone is placed in front of an open refrigerator which has beer in it, then he will believe that there is beer in the refrigerator.
>
> If someone wants beer and believes that there is beer in the refrigerator, then he will go to the refrigerator and get a beer.
>
> If someone believes that there is beer in the refrigerator, then he believes that there is beer somewhere.
>
> If someone believes that there is a Budweiser in the refrigerator, then he believes that there is beer in the refrigerator.
>
> If someone intends to get a beer out of the refrigerator, then he believes that there is a beer in the refrigerator.

As in the detective story, let's introduce a label for the belief that there is beer in the refrigerator; let's call it 'state X':

> If someone is placed in front of an open refrigerator which has beer in it, then he will be in state X.
>
> If someone wants beer and is in state X, then he will go to the refrigerator and get a beer.
>
> If someone is in state X, then he believes that there is beer somewhere.

> If someone believes that there is a Budweiser in the refrigerator, then he is in state X.
>
> If someone intends to get a beer out of the refrigerator, then he is in state X.

As in the case of the detective story, corresponding to this label 'state X' is a certain *role* in the story: 'state X' stands for whatever state one is in when one is in front of an open refrigerator which has beer in it, *and* which, together with the desire for beer, causes one to go to the refrigerator and get a beer, *and* which one is in when one believes that there is a Budweiser in the fridge ....

In the case of the detective story, we said that a person could realize one of the roles in the story if they did all of the things which the role included. Just so, in this case, we can say that an *internal state of a person* can realize the 'state X' role in our story if it does all of the things included in the role.

What is the property of believing that there is beer in the fridge? It is the property of having a state which realizes the role laid out in claims like the ones above. (Of course, a realistic story will contain many more claims.)

How might this work for the property of having a toothache?

If you understand all of that, then you can understand functionalism. Functionalism is the idea that we can tell a (much longer) story like this for every mental property. Each one of these stories defines, corresponding to each mental property, a certain *role* (sometimes called a *functional role*). What it is for someone to have that mental property, according to the functionalist, is for them to have some state which realizes that role.

Let's consider a few objections to functionalism, as stated.

Objection 1: the theory is circular, since it defines one mental property in terms of other mental properties.

This is correct. Really, to explain the functional role with which a mental state is identified, we would have to replace every occurrence of a term for a mental state with a 'label.' So our example would really be more like

> If someone is placed in front of an open refrigerator which has beer in it, then he will be in state X.
>
> If someone is in state Y and is in state X, then he will go to the refrigerator and get a beer.
>
> If someone is in state X, then he is in state Z.
>
> If someone is in state A, then he is in state X.
>
> If someone is in state B, then he is in state X.

This makes it more like the example of the detective story, in which we have multiple characters.

Does the view look less plausible when we make it non-circular?

Objection 2: no person has states which exactly fit the above functional role. After all, it is not always the case that when I desire a beer and believe that there is a beer in the fridge, I will go to the fridge and get the beer.

This is why Putnam characterizes his theory in terms of a probabilistic, rather than a deterministic, automaton. If we think that psychological laws will always be merely probabilistic, it seems that any plausible functionalist theory will have to be qualified to respect this fact.

Objection 3: Putnam's super-super-spartans don't have any state which realizes the functional role of pain, since they have no state which typically causes them to wince, cry out, etc. But, despite this, they feel pain. So the same examples which show that behaviorism is false also show that functionalism is false.

How would Putnam respond to this objection?

3 Is functionalism a form of materialism?

The kind of functionalist theory we have been talking about seems very different than Descartes' view of mental properties. But Putnam says that his theory is not incompatible with dualism. What's going on?

Let's suppose that we have some description of the functional role of pain – call this the *pain role*. Then Putnam's theory is:

> The property of being in pain = the property of having some state which plays the pain role.

'State' here seems to mean just 'property.' So we could restate the view as

> The property of being in pain = the property of having some property which plays the pain role.

Playing the pain role is what is sometimes called a higher-order property: it is a property of properties. In my case, a property of neurons might have this higher-order property; in the case of an alien, it might be some property of a bunch of silicone; in the case of an angel, it might be some non-physical property. So far as the functionalist is concerned, which property plays this role does not matter — to feel pain is just to have some property or other which plays this role.

So it is unlike the identity theory in that it is compatible with substance dualism; but it is unlike Descartes' view in that it is compatible with materialism.

How does functionalism fare with respect to the problems we have seen for the other three views of mental properties we have discussed?