8. More on machine functionalism

Martín Abreu Zavaleta

June 6, 2014

1 Functional properties and realizations

Functionalism is the view according to which to be capable of mentality is to have a certain functional organization. Let's use a simple functional definition of the state of being in pain to illustrate:

> Being in pain = being in a state such that sitting on a tack causes that state and being on that state causes anxiety and saying 'ouch'.

Using this simple definition, we can make a distinction between a *functional role*, a *realizer* and a *functional property*:

Functional role: A functional role is the property of standing in such and such relations to other properties, states, etc., as determined by a given description. In the example above, the functional role associated with pain is this: the property of *being the state* that is caused by sitting on a tack and causes anxiety and saying 'ouch'. Note that the property of being *that* state is not the same as the property of being *in that state*. A person can be in that state, but it wouldn't make much sense to say that the property *is* that state.

Realizer: A realizer is a property, state, etc. that plays a certain functional role. For instance, if stimulation of C-fibers is a state that is caused by sitting on a tack and causes anxiety and saying 'ouch', then stimulation of C-fibers *realizes* the pain-role as defined above.[1]

Functional property: A functional property is the property of having a property of that realizes a certain functional role. For instance, the property of being in pain is the property of having a property that realizes the pain-role. In this case, one of the ways in which one can have the property of being in pain is by one's C-fibers being stimulated.

We will discuss these definitions more thoroughly when we examine Lewis's paper, but it's a good idea to get used to them and keep them in mind. Now let's go back to Putnam's machine functionalism.

[1] Another way of saying this is that the stimulation of C-fibers realizes the functional role determined by the definition of pain above.

2 Identifying states

On Putnam's view, the required functional organization is given by a machine table. Thus, on this kind of functionalism, to be capable of mentality is to have a functional organization *describable* by a certain machine table.

With respect to particular mental states, functionalism tells us that to be in a mental state *m* is to be in a state that plays the *m-role*, as defined by the appropriate machine table. Note, however, that machine tables define states in relation to all the other states, inputs and outputs in the table: it doesn't make much sense to ask whether the state called 'A' in one machine table is the same as the state called 'A' in a different machine table. The reason is that states are defined with respect to the rest of the states in a single machine table.

But this means that in order for something to be in a particular mental state, it has to have exactly the structure characterized by a given machine table. So any two things that can have that mental state must have the same psychology! Kim puts the point nicely:

> the machine-functionalist conception of the mind has the following consequence: For any two subjects to be in the same mental state, they must realize the same Turing machine. But if they realize the same Turing machine, their total psychology must be identical. That is, on machine functionalism, two subjects' total psychology must be identical if they are to share even a single psychological state—or even to give meaning to the talk of their being, or not being, in the same psychological state. This sounds absurd: It does not seem reasonable to require that for two persons to share a mental state—say, the belief that snow is white—the total set of psychological regularities governing their behavior must be exactly identical. (Kim, p. 152)

So, on one hand, functionalism seems to be compatible with the multiple physical realizability of mental states. On the other, it seems to require that everyone's total psychology be exactly the same. This doesn't sound promising, but perhaps it can be solved by requiring that partial characterizations of mental states be given, but it's unclear whether this can be done.

Another problem comes when defining the inputs and outputs for the functional roles of mental states. Maybe humans say 'ouch' when they have pain, but octopuses don't. So we need a characterization that is general enough to cover all cases that we are interested in. Again, it's not clear what kind of story we should give about this, and we won't go too deep into the details.

These objections can be seen as objections of detail: it is possible that, in the future, we will come out with a way of refining the theory that solves these problems. However, other people have raised stronger objections to functionalism, or tried to offer arguments that restrict the extent of its success. We'll start with Searle's famous *chinese room argument*. Then we'll see Lewis's version of functionalism, and see more important objections by Ned Block.

3 Searle's Chinese room argument

Searle invites us to consider a case like the following:

> There's a person in a room who doesn't speak chinese, but can distinguish symbols by their shape. She is put in a room and given two pieces of paper with chinese symbols in it. She is also given a book with instructions (in English) for how to manipulate chinese

> symbols. the book gives her instructions to write, for any combination of symbols in the two pieces of paper that she is given, a new combination of symbols in a different piece of paper. Every time she gets the papers with the chinese symbols in them, she follows the instructions in the books, writes some symbols in a new piece of paper, and gives this to someone through a little window. This results in what appear to be the responses that a native chinese speaker would give to a series of questions.

Searle thinks that if functionalism was true, then the person in the room would understand chinese merely by virtue of following the rules in the book, but he thinks that the person in the room clearly doesn't understand chinese.

Notice that the example is not especially directed against a functionalist characterization of *understanding chinese*. Any other mental state could have helped, as long as we take the instructions in the book to give a complete functional description of the mental state in question, and the processes executed in the chinese room are the same functional processes as the ones executed when someone has the relevant mental state.

What should we think about this argument? Searle considers a couple of objections and replies, the most important of which is the **system reply.** According to the system reply, while it might be true that the person inside the room doesn't understand chinese, what matters is that the whole system does; that is, that the system constituted by the person in the room, the book of instructions, the pieces of paper, and the processes by which the person manipulates the symbols, *does* understand chinese.

Here is Searle's response:

> My response to the systems theory is quite simple: let the individual internalize all of these elements of the system. He memorizes the rules in the ledger and the data banks of Chinese symbols, and he does all the calculations in his head. The individual then incorporates the entire system. There isn't anything at all to the system that he does not encompass. We can even get rid of the room and suppose he works outdoors. All the same, he understands nothing of the Chinese, and a fortiori neither does the system, because there isn't anything in the system that isn't in him. If he doesn't understand, then there is no way that the system could understand because the system is just a part of him.

Let's focus first on one of the things he says: there isn't anything in the system that is not in the operator, and so, Searle says, if the operator doesn't understand Chinese, neither does the system. Is this a good argument?

No. Consider the following lines of reasoning:

1. There isn't anything in my heart that is not in me, so if I don't weigh five pounds, neither does my heart.

2. There isn't anything in the Chinese room that isn't in the operator, so if the operator wasn't designed by a group of programmers, neither was the Chinese room.

These are all bad arguments, so why should it be good when replying to the systems response?

A more important point is the following: just because the operator doesn't understand Chinese, that doesn't mean that the system "inside" him doesn't understand either. Think about *virtual machines*: it is possible to have a given operative system, say Mac OS, and have it emulate another

operative system, say some version of Linux. Now it is possible that the Linux OS is running a certain program, say Firefox, but this doesn't by itself mean that MacOS is also running Firefox: MacOS may not even have Firefox installed in it!

It may even be that the Linux machine crashes while MacOS, and even the emulator itself, is still running. Moreover, there is some sense in which Linux is "inside" or "incorporated" into MacOS, but as we've seen, this doesn't mean that they have to share their states.

This is the kind of relation between the operator and the Chinese room inside her head: the fact that the operator has internalized all the process and instructions doesn't mean that the operator must share all of the Chinese room's states. For instance, the Chinese room may believe that China was at its cultural peak during the Han dynasty, but this doesn't entail that the operator must also be in that state, and vice versa.