Philosophy 355: Contemporary Philosophy — Hamilton College
Fall 2008 — Russell Marcus
Tuesdays and Thursdays, 9am - 10:15am — rmarcus1@hamilton.edu

Class 5 - Troubles with Functionalism

I. The biological versus the artificial

Jackie, in class on Tuesday, worried about the presumption that we are making, at least in the first example of multiple realizability, that machines can have mental states.
Adeline wondered if the difference between the machines and human beings has some biological basis.
John Searle, famously, constructed a thought experiment, called the Chinese Room, and concluded from it that there is something essentially biological about mentality.
Searle was responding to Putnam's claim that we can test functionalism by constructing models of human minds.
To understand minds, according to Putnam's functionalist, we can examine computer models and their software.
Computers and their software work according to purely formal, syntactic manipulation.
They merely follow algorithms, every step of which can be specified syntactically.

Searle's Chinese room example provides an example of a person working according to purely formal, syntactic rules.

> Suppose that I'm locked in a room and given a large batch of Chinese writing. Suppose furthermore (as is indeed the case) that I know no Chinese, either written or spoken, and that I'm not even confident that I could recognize Chinese writing as Chinese writing distinct from, say, Japanese writing or meaningless squiggles. To me, Chinese writing is just so many meaningless squiggles. Now suppose further that after this first batch of Chinese writing I am given a second batch of Chinese script together with a set of rules for correlating the second batch with the first batch. The rules are in English, and I understand these rules as well as any other native speaker of English. They enable me to correlate one set of formal symbols with another set of formal symbols, and all that 'formal' means here is that I can identify the symbols entirely by their shapes. Now suppose also that I am given a third batch of Chinese symbols together with some instructions, again in English, that enable me to correlate elements of this third batch with the first two batches, and these rules instruct me how to give back certain Chinese symbols with certain sorts of shapes in response to certain sorts of shapes given me in the third batch. Unknown to me, the people who are giving me all of these symbols call the first batch a script, they call the second batch a story, and they call the third batch questions. Furthermore, they call the symbols I give them back in response to the third batch answers to the questions, and the set of rules in English that they gave me, they call the program. Now just to complicate the story a little, imagine that these people also give me stories in English, which I understand, and they then ask me questions in English about these stories, and I give them back answers in English. Suppose also that after a while I get so good at following the instructions for manipulating the Chinese symbols and the programmers get so good at writing the programs that from the external point of view - that is, from the point of view of somebody outside the room in which I am locked - my answers to the questions are absolutely indistinguishable from those of native Chinese speakers. Nobody just looking at my answers can tell that I don't speak a word of Chinese. Let us also suppose that my answers to the English questions are, as they no doubt would be, indistinguishable from those of other native English speakers, for the simple reason that I am a native English speaker. From the external point of view - from the point of view of

> someone reading my "answers" - the answers to the Chinese questions and the English questions are equally good. But in the Chinese case, unlike the English case, I produce the answers by manipulating uninterpreted formal symbols. As far as the Chinese is concerned, I simply behave like a computer; I perform computational operations on formally specified elements. For the purposes of the Chinese, I am simply an instantiation of the computer program. (From John R. Searle, "Minds, Brains, and Programs")

The person in the Chinese room has all the same input as a speaker of Chinese, and produces the same output, without having any understanding of Chinese.

Even if he internalizes all the formal rules, Searle in the Chinese room lacks any understanding about the content of the symbols he is manipulating.

Searle's argument is as follow:

1. Programs are completely describable in terms of their formal, syntactic content.
2. Minds grasp the meanings, or semantics, as well as syntax.
3. Syntax alone can not produce semantics.

So, minds are not merely syntactic manipulators; i.e. minds are not mere programs.

The importance of Searle's argument, for functionalism, is that a mechanical model of the mind could not *be* a mind.
There seems to be more to our minds than algorithmic processing of sensory input.
The functionalist was motivated by the desire to account for multiple realizability.
Searle's argument is that there is a virtue in chauvinism, that there is something essentially biological about our intentionality.
But, we will not spend a lot of time on Searle's argument.
Searle's argument focuses on intentionality, rather than consciousness, which is our main quarry.
And, Searle's argument, while it may be taken as an objection to functionalism, is aimed more directly at the question of whether there can be artificial intelligence.
Here, we are accepting that there could, in principle, be artificial intelligence.
We want a definition of mind that will accommodate it.
Further, even if we are in principle committed to the claim that machines could not have mental states, we might be willing to consider silicon- (or other element-) based aliens as mental creatures.
And even further, there are other problems of multiple realizability, like neurological equipotentiality and non-relational construals of mental states, which motivate functionalism away from identity theory.

II. Block's definitions of functionalism and the Ramsey sentence formulation

Block differentiates between machine functionalism and non-machine functionalism, and also between Functionalism and Psychofunctionalism.

He defines machine functionalism in terms of Turing machines.
He defines non-machine functionalism in terms of Ramsey sentences.

(I worry that these definitions amount to the same thing, in the end.
Perhaps there is a difference that I do not yet grasp.

Putnam identifies mental states with the states of the machine table.
But, there is an interesting problem with machine functionalism.
Our mental states can not correlate with machine-table states because there are more mental states that humans can have than there are machine-table states.
So, the functionalist should want not to be a machine functionalist.
But, I'm not sure how to avoid the problem, except by fiat.
In any case, nothing we do here will depend on the difference between machine and non-machine functionalism.)

The definition of functionalism in terms of Ramsey sentences, especially the technical discussion in the long paragraph on pp 272-3, is worth examining in more depth than I provided in the last class notes.

Consider a psychological theory T.
The theory is just a long set of sentences correlating inputs, outputs, and mental states.
Remember, that we took functionalism to be a lot like behaviorism with the inclusion of internal states.

$T(s_1...s_n, i_1...i_m, o_1...o_k)$

T contains three kinds of terms:

The 'i's are terms for inputs.
The 'o's are terms for outputs.
The 's's are terms for mental states.

For example, T might include terms for seeing a cylindrical patch of orange; desiring an orange soda; enjoying an orange soda, and saying, 'Ahh, I enjoyed that orange soda'.

i_{7345} = having an orange soda can in front of you
s_{2342} = seeing the cylindrical orange patch
s_{4873} = desiring orange soda
s_{92357} = enjoying an orange soda
o_{983} = Saying, 'Ahh, I enjoyed that orange soda'

T might thus say (that is, it might be a theorem derivable from T) that whenever a person is in state s_{4873} and receives input i_{7345} so that she moves develops state s_{2342}, she also moves into state s_{92357} and produces output o_{983}.
T entails lots and lots of these theorems, for every different mental state, and combination of mental states.

The behaviorist theory, call it B, would look a lot like T.
B would have terms for inputs and outputs, the $i_1...i_n$, and the $o_1...o_n$.
But, in contrast to T, B would require the elimination of internal states.

$B(i_1...i_n, o_1...o_m)$

That is, the behaviorist just correlates inputs and outputs.
We could introduce terms for mental states as shorthand for some subset of the correlations in B.
But mental state terms would not be used in the austere version of the theory.

Thus, behaviorism is more parsimonious than both dualism and identity theory.
But, the theorems of B would be more difficult, or impossible, to develop, since references to mental states could not be used to differentiate, say, honestly expressing one's enjoyment and lying about it.

The identity theorist's psychological theory, would require reference to brain (or neural) states.
The $s_1...s_n$ in T referred to mental states, like seeing an orange patch, or feeling pain, or believing that snow is white.
For the identity theorist, these terms would have to refer to particular brain states, or, perhaps, brain and body states.
The phenomena of multiple realizability were supposed to show that such a theory was unlikely.
We can conclude that any theory of mind must satisfy a multiple realizability condition, that terms for mental states have to be able to refer to different kinds of physical states.
(By analogy, we might take 'water' to refer to both H_2O and XYZ.)
The dualist, note, can satisfy the multiple realizability condition by reifying the mental states, and then just correlating them with a variety of physical states.
So, how does the functionalist satisfy the multiple realizability condition.

Let's go back to T, the psychological theory, and its terms for mental states, the $s_1...s_n$.
The identity theorist says that these terms refer to brain (or brain and body) states.
The functionalist can not have the $s_1...s_n$ refer to any kinds of physical states, though, since any physical referent would violate the multiple realizability condition.
If the functionalist adopts dualism, they can have them refer to states of an immaterial soul, but most functionalists are token physicalists, so the dualist option is unacceptable.
Instead, the functionalist chooses to deny that they refer to any thing at all!
The functionalist replace the singular terms with variables, and then quantifies over them to form the Ramsey sentence of the theory.

The following analogy for Ramsey sentences comes from David Lewis, "Psychophysical and Theoretical Identifications".

> 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. Call these theoretical terms (T-terms for short) because they are introduced by a theory.

The names in the story are analogous to the names of our mental states.
We need not know anything about them, for the theory to make sense, except that they are some kinds of things with the relations that the theory posits, among the other elements: the inputs, the outputs, and the other mental states.
So, the mental state of enjoying an orange soda is whatever is caused by the inputs and other mental states that the theory claims precede it, and which causes the mental states and output that the theory claims it produces.
The mental state is whatever plays the causal role of that state, in the psychological theory.
So, now, the functionalist theory looks like this:

$$\exists x_1 ... \exists x_n \, T(x_1 ... x_n, i_1 ... i_m, o_1 ... o_k)$$

To say that a person is an a particular mental state, like the state of enjoying an orange soda, we can ignore Block's property abstraction operator.

$$\text{p is enjoying an orange soda iff } \exists x_1 ... \exists x_n \, T(x_1 ... x_n, i_1 ... i_m, o_1 ... o_k \text{ and p is in } x_{92357})$$

The functionalist meets Fodor's demand that our psychological theory have a relational construal of mental states by defining the mental states by its relations to other mental states and inputs and outputs.
The chauvinism of the identity theory, which claims that each of the singular terms of T reduce to brain states, is avoided, since we do not look to reduce these terms to particular physical states.

(And, here again, is my worry that there is no difference between machine and non-machine functionalism.
It looks to me that the (non-machine functionalist's) theory T is just a Description of the Turing machine to which Putnam's machine functionalist ascribes mentality.)

III. Inverted and absent qualia

Block calls his central argument against functionalism the Absent Qualia Argument.
It is related to the problem of inverted qualia, which appears originally in Locke's *Essay*.
Locke's idea was that two people could be identical in their behavior, and indeed in their functioning, and yet not share the same phenomenal experience.

One version of the problem arises from mere differences in physiology.
My eyes are perhaps a bit bigger or smaller than yours.
Perhaps you have more rods or cones, which are the physical basis for color perception.
Why should I believe that my sensation of red matches yours?
In fact, since I am color blind, we have no reason to believe that we have the same perceptions.
But, we don't share the same functions, either.
The problem arises for two people who do see all colors.
One person's experience might be more vibrant, or brighter, or slightly shifted to the left.

The more startling problem arises from considering two normal sighted people who agree on a whole range of color ascriptions.
What if every time one saw red, the other saw purple; every time one say blue, the other saw green?
They could still use the same terms; they would be functionally isomorphic.

But, they would be having different qualia.
The problem for functionalism is that if there are cases of inverted qualia, then people with the same functional states are in different mental states.
And, there seems to be no way to deny the possibility of inverted qualia.
So, functionalism fails to capture the nature of our mental states.
Notice, as Block does, that the idea of inverted intentions is almost impossible to formulate, 288.
So, Block argues, qualia supervene on functional organization.

Block's Absent Qualia Argument goes one step further.
The brain is essentially a collection of neurons, which discharge impulses from one to another.
Neurons fire, and induce other neurons around them either to fire or not to fire.
The story is more complicated, of course, but the differences appear only to be a matter of degree, not of kind.
The basic picture of neurons transmitting information like electrons passing along a circuit board is apt.
Block presents several different thought experiments to show that there are functional equivalents of minds which lack properties that minds should have.
Thus, where identity theory was too chauvinist, functionalism is too liberal, ascribing minds to too many things.

One problem of liberalness comes, late in the article, in the specification of the $i_1...i_n$s and the $o_1...o_m$s.
On the one hand, we want to avoid chauvinism by specifying inputs and outputs in terms of human physiology, like neural stimulations, or sense receptors.
But, Block argues convincingly, it will be difficult to avoid chauvinism this way.
On the other hand, if we just Ramsify the inputs and outputs, we seem to be able to count anything as a system, p 294.
But, the central problem is not about specifying the inputs and outputs, it is about the missing qualia.

In the homunculi-headed robot examples, the brain of a creature functionally equivalent to me turns out to have tiny persons inside his brain, rather than neurons, performing exactly the same functions that the neurons perform in my head.
In the Chinese nation example, Block imagines that we have mapped the brain, and it contains one billion neurons.
(This is a fiction, but only by a factor of about a hundred - there are about a hundred billion neurons in the brain.)
Now, we can set up the people of China to act as this billion-neuron brain, with walkie-talkies and connecting each person to surrounding people.
We give each person the instructions to transmit information in the way that our neurons do, to other people.
The brain can be attached to a human sensory organs via radio signals from the receptor nerves.
That is, we would connect a creature that looked and functioned just like us with an artificial processing system made out of the people in China.

In both inverted and absent qualia cases, the functionalist seems to fail to account for occurrent sensory states.
One possible response for the functionalist is to argue that the systems Block describes do not actually lack qualia. Discuss.

Furthermore, what do we think about the space-ship molecule people, pp 279-80?

IV. Functionalism and Psychofunctionalism

Another possible response is to denigrate the importance of qualia.
Consider Block's discussion of Dennett, 290-1.
(We'll get back to Dennett, later.)

Block distinguishes between two forms of functionalism.
Functionalism allows ordinary terms referring to mental states for the $s_1...s_n$.
Psychofunctionalism may avoid some common-sense terms, and even allow terms for specific neural states, instead.
The key difference is that Psychofunctionalism defines and individuates mental states according to our best scientific psychological theory, p 272.

Ex hypothesi, the homunculi-headed (or Chinese-nation-headed) robots behave just like ordinary people.
We might want to attribute beliefs, desires, and other intentions to the robot.
Indeed, Block argues that the Psychofunctionalist would have to do so, 285-6.

If we can agree with the Psychofunctionalist that we should attribute intentional states to his functional equivalents, then if the Psychofunctionalist can argue that qualia have no room in our best scientific psychological theory, perhaps functionalism can stand, p 289.
Block presumes that we must account for qualia, and so the argument that there is no room for them in scientific psychology reinforces his argument that Psychofunctionalism is untenable.
The Dennett approach, which we will examine soon, is to deny the reality of qualia at all.

V. A final worry about functionalism

A last worry about functionalism, due to Fodor and Block, concerns dispositional states, and simultaneous occurrent states.
An occurrent state is one that can happen at a particular time.
So, if I am seeing a blue shirt, I am having an occurrent sensation.
The functionalist identifies this sensation with its causal relations with other color sensations.
If there is a state of the program which corresponds to the seeing of blue, then the instantiation of the program can be said to be in that state.
If I see a tiger in the room, I would have an occurrent fear.
Again, the program can represent that occurrent fear.
But, if I am just afraid of lions, that is a dispositional state.
Similarly, it seems pretty clear that you believe that 1476+1=1477.
You believed that, even before I mentioned it.
But, you weren't thinking about it.
It was just a disposition.

Functionalism has a good explanation of the causal relations among sequential mental states, as the causal relations among the steps in the machine table/program.
But, how is the functionalist supposed to represent simultaneous, but distinct mental states?
I might at the same time, see the tiger in the room, fear the tiger, be excited about the tiger, etc.
These are distinct mental states, but there is only one state of the program at any one time.
The functionalist thus lacks an explanation of the interactions among simultaneous mental states.

To accommodate dispositions and simultaneous occurrent states, the functionalist can appeal to the fact that the machine table has the potential to be in a particular state.
The occurrent state can be identified with the particular state of the machine instantiating the program.
The dispositional state can be identified with a broader swath of the machine table.
But, what corresponds to the disposition is the whole machine table, not a particular state of it.
Thus, our dispositional states will be any states that the program can be in, which correlates with anything that we can possibly think.
But, it is possible for us to think all sort of things that we do not believe.
It is possible for the program to be in lots of states that do not correspond to actual beliefs.
We could pick out some of the states as corresponding to our beliefs.
But then, we can not say that two organisms are in the same mental state if and only if they have the same states of their tables.
This seems like giving up on functionalism.