Class 21 - April 13
Functionalism

I. Functionalism

In our last class, we noticed that the materialist theory of mind suffered from some problems of multiple realizability.
By typing (defining) mental states according to their neurological correlates, we produce a chauvinist theory that fails to grasp essential aspects of our mental lives.
Relatedly, type physicalism may provide a good third-person account of mental states, but misses their first-person qualities.

Functionalism was designed to avoid the problems we have seen, but maintain the useful insights of both behaviorism and identity theory.
The functionalist describes the mind by appealing to an analogy with computer software: the mind is the software of the brain.
Just as the same software can be run on different hardware, the same mental states can be instantiated by distinct physical (or, even, non-physical) systems.
The mind is not identified with any particular hardware.

> In the functionalist view the psychology of a system depends not on the stuff is it made of (living cells, metal or spiritual energy) but on how the stuff is put together (Fodor 451).

Two things are in the same mental states if, and only if, they have the same state of their programs.
Being in pain, or seeing blue, or believing that the moon is made of cheese, are functional states of an organism.

Functionalism is often presented as the claim that minds are Turing machines with sensory inputs and motor outputs..
The effect of an input can be merely to change the state of the system, with no motor output.
Or, an input can lead directly to an output.
Turing machines are computers, basic information processors.
A Turing machine contains, in its machine table, a complete list of possible states of the system, and possible inputs, and the output.
The actions of a Turing machine (what it writes, where it goes, what state it is in) are completely determined by its algorithm, or set of rules.

An algorithm is a list of instructions, a procedure.
Computer programs are algorithms; cooking recipes are algorithms.
Recipes generally just give simple, linear instructions.
An algorithm can also do different things depending on the state of the system executing the algorithm.
Thus, some algorithms, like the one we generally use for long division, contain conditional clauses: if...then... statements.

Essentially, a computer is just a mechanism which reads input, has internal states, and computes output on the basis of those states and its instructions.
Every action the computer takes is completely determined by the algorithm governing its states and its input.

Fodor uses an analogy of a vending machine to explain the functionalist theory of mind.
Vending machines are defined by the ways in which they return output (say, snacks) on the basis of their inputs (money and pushed buttons for selections).
When you put money into a vending machine, it changes the internal states of the machine without changing any motor output.
It can be in a state such that it won't give you anything you select.
But, eventually, if you give it enough money and make a selection, it will dispense something and perhaps return change.
Further, the program of the vending machine is independent of the particulars of the machine.

> Nothing about the way I have described the... Coke machines puts constraints on what they could be made of. Any system whose states bore the proper relations to inputs, outputs and other states could be one of these machines. No doubt it is reasonable to expect such a system to be constructed out of such things as wheels, levers and diodes (token physicalism for Coke machines). Similarly, it is reasonable to expect that our minds may prove to be neurophysiological (token physicalism for human beings) (Fodor 456b).

One difference between humans and vending machines is that very few people think that vending machines have free will.
Vending machines pretty much always do what we make them to do.
When they don't their deviations from the norm are explicable in accordance with physical laws.
Organisms with mental states, on the other hand, do not appear move from one state to the next with certainty.
It might be that the world is strictly deterministic, but it might not.
Since we do not know whether our actions are completely determined, it would be nice to have a theory of mind which could accommodate freedom.
One version of functionalism which does so, due to Hilary Putnam, says that mental states are functional states of probabilistic automata.
A probabilistic automaton also has sensory inputs and motor outputs, just like a standard Turing machine.
It has the same structure as a completely deterministic machine, but with probabilistic responses.
It is not clear that our freedom is the same as being a probabilistic automaton, but they are both non-deterministic.

II. The Birth of Functionalism

Functionalism attempts to integrate some of the salvageable aspects of behaviorism, most notably its reliance on the relations among sensory input and behavior/mental states.
According to both the behaviorist and the functionalist, we type mental states according to behavior, not according to the qualities available by introspection.
I called such typing, following Fodor, a relational construal of mental states.
The functionalist takes behaviorism's attributions of mental states based on behaviors, and removes its disavowal of internal states, and its reductionist, eliminativist, program.
Behaviorism tried to reduce mental state language to behavior language, with the goal of eliminating any apparent references to immaterial substance and internal causes.

> The radical behaviorist predicts that as psychologists come to understand more about the relations between stimuli and responses they will find it increasingly possible to explain behavior without postulating mental causes. The strongest argument against behaviorism is that psychology has

> not turned out this way; the opposite has happened. As psychology has matured, the framework of mental states and processes that is apparently needed to account for experimental observations has grown all the more elaborate. Particularly in the case of human behavior psychological theories satisfying the methodological tenets of radical behaviorism have proved largely sterile, as would be expected if the postulated mental processes are real and causally effective (Fodor 452).

Functionalism, in contrast to behaviorism, is compatible even with substance dualism, since it makes no claim about where and how mental properties are instantiated.

In parallel fashion, the functionalist adopts from identity theory the legitimacy of mental states and an acceptance of the causal connections among them.

> Since mental states interact in generating behavior, it will be necessary to find a construal of psychological explanations that posits mental processes: causal sequences of mental events (Fodor 454a).

The functionalist dispenses with identity theory's unacceptable chauvinism.

> The problem with type physicalism is that there are possible information-processing systems with the same psychological constitution as human beings but not the same physical organization. In principle all kinds of physically different things could have human software (Fodor 455a)

But functionalism adopts identity theory's claim that there are explanations which refer to mental causes.

> Functionalism construes the concept of causal role in such a way that a mental state can be defined by its causal relations to other mental states (Fodor 455b).

III. The Turing Test

Turing machines are named after Alan Turing, an early computer scientist and defender of artificial intelligence.
Turing argued that we could replace the question of whether machines can think with the question of whether computers could fool people into believing they were human.
This strategy for determining whether a machine is thinking is called a Turing test.

One might believe that to determine whether something has thoughts, one must find out whether it has conscious mental states.
The problem with Cartesian criteria is that they are impossible to apply.
Turing says that the Cartesian criterion for thought (consciousness) leads to a solipsistic point of view.
We can not successfully apply the Cartesian criteria even to our closest friends and family.
So, we can never know that other people have mental states, either.
This is the problem of other minds.
Turing's test for whether a machine thinks denigrates any Cartesian criteria and relies on behavioral criteria for ascriptions of thought.

Since computers don't have brains, the identity theorist says that they can not think.
As we saw, that claim appears chauvinist.

If it turns out that minds are essentially Turing machines, it will follow that machines can think.
So for Turing, the question of whether a machine can think is precisely a question about what the machine can do.

There seems to be some room between Turing's position, that anything that acts like a person thinks, and the solipsistic Cartesian position, that the only things to which we can attribute thought are ones which we can verify are actually conscious.
We might be able to find empirical criteria for establishing consciousness which could moderate the Cartesian view.
Or, we could adopt a view about consciousness which is tied neither merely to behavior or to only our introspective mental states.

IV. Ramsey Sentences

Our central concerns about identity theory were collected under the umbrella of multiple realizability.
The functionalist avoids problems of multiple realizability by identifying each mental state with the relevant properties of that state, its interactions with other mental states and the behaviors of people in that mental state.
Functionalism includes no reference to irrelevant particulars like brain states.
A thing is in pain if it has been affected in certain relevant ways, and if it has other concomitant mental and behavioral states (like wincing or crying) which are causally related to it.
Nothing about the particular material in which the causal processes happen is relevant to the definition of the mental states.
When we are researching the mental states of a particular organism, we will of course look at the specific causal processes involved.
But, when we generalize to a functionalist theory of mind, we abstract away from physical particulars.

Functionalists rely on a logical trick to eliminate irrelevant vocabulary from the theoretical identity sentences of a formal theory of mental states, to achieve the desired level of abstraction.
We can start with a chauvinist physicalist theory of the mind and turn it into an acceptable functionalist theory.
To do so, we can construct what are called Ramsey sentences.
Ramsey sentences replace specific references to particular causal structures (say, brain states) with claims that something (anything) has this causal role.
Imagine a description of your whole life: your experiences, your various mental states and how they are connected, the (presumably causal) relationship between your body, including your brain, and those mental states, the resulting behavior.
To Ramsify this description, replace references to the specifically mental parts of this theory, pains and color terms and beliefs, with variables.

Let's see in more detail how such a Ramsification would proceed.
Consider a psychological theory T.
T is just a long set of sentences correlating inputs, outputs, and mental states.
Remember that we took functionalism to be 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 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 mental state and combination of mental states and inputs and outputs.

A 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 $o_1...o_n$.
In contrast to T, B would contain no references to internal states.

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

That is, the behaviorist tries to provide a satisfying psychological theory by just correlating 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.
Our main criticism of behaviorism could be expressed by saying that the theorems of B would be impossible to develop since references to mental states could not be used to differentiate honest expressions of one’s mental states from fakery.

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.
The dualist, note, can satisfy the multiple realizability condition by reifying the mental states and correlating them with a variety of physical states.

The functionalist satisfies the multiple realizability condition by claiming that the $s_1...s_n$ can refer to any kinds of states,

Any particular physical references would violate the multiple realizability condition.
The functionalist can even claim that the $s_1...s_n$ refer to states of an immaterial soul.
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.

In the terms of first-order logic, the functionalist is existentially quantifying over the $s_1...s_n$.
Now, the functionalist theory looks like this:

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

We can define a person's particular mental state, like the state of enjoying an orange soda, thus:

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

The resulting theory provides a functional, causal-role definition of your mental states.
It defines mental states in terms of their functional roles.
The functionalist can characterize some mental states, like pain, even more generally.

> x is in pain iff x has been affected by the kinds of things that cause pain, has other mental states that generally accompany pain, and exhibits the kind of behavior that are associated with pain.

Pain is whatever has the place that pain has in your life.
It is preceded by physical or emotional blows, and succeeded by characteristic behavior: sometimes avoidance, and sometimes valiant confrontation.
It engenders certain other mental states, fear or anger or resignation, all of which have their own causal-role definitions.
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.

The functionalist meets the demand that our psychological theory have a relational construal of mental states by defining mental states by their 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.

V. The Biological and the Artificial

The functionalist is naturally aligned with the defender of artificial intelligence.
If mental states are the kinds of things that do not depend on any particular kind of material instantiation, then there is no presumption that any one kind of material basis for thought is privileged.
Against this liberal view of mind, we might wonder if the difference between machines and human beings has some biological basis.
John Searle famously constructed a thought experiment, called the Chinese Room, to argue that there is something essentially biological about mentality.
Searle was responding both to claims of machine intelligence and to claims that we can test functionalism by constructing models of human minds.
To understand minds, according to the 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 (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.
But she has no understanding of Chinese.
Even if she internalizes all the formal rules governing the correlation of questions and answers, she lacks any understanding about the contents of the symbols she 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.

Searle's argument is generally presented as an argument against artificial intelligence.
Any artifact, even a very complex machine, would have to be merely faking it.
The importance of Searle's argument for functionalism is that it seems to show that a mechanical model of the mind could not *be* a mind.
If we take the computer analogy seriously, the functionalist says that the mind is the software, the structure or organizing principles, the brain.
Minds are essentially just information processing.
But, the Chinese room example seems to show that there is more to our minds than algorithmic processing of sensory input toward the production of motor output.

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.
We will not spend more time on Searle's argument.
We might accept that there could be artificial intelligence and want a definition of mind that will accommodate it.
Or, even if we are in principle committed to the claim that machines could not have mental states, we might be willing to consider other biological organisms as mental creatures (e.g. chimps, dolphins).

Even further, there are other problems of multiple realizability, like neurological equipotentiality and the non-relational construal of mental states, which motivate functionalism away from identity theory.

VI. Inverted and Absent Qualia

One of the problems we saw with prior materialist theories of mind is that they fail to capture our internal mental lives.
The behaviorist rejected all first-person evidence as misleading and useless.
The identity theorist accepted that internal states were causes of behavior, rather than identical to behavior, but identified the mental states with their neural correlates.
Both theories work better as third-person accounts of mental states than as first-person accounts.
Similarly, functionalism has been criticized for failing to account for the way that mental states appear to us, for consciousness.

One consciousness-related argument against functionalism is called 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 for phenomenal or qualitative states arises from mere differences in physiology.
My eyes are perhaps a bit bigger or smaller than yours.
Maybe 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 good reason to believe that we do not 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 of the inverted spectrum arises from considering two normal-sighted people who agree on color ascriptions.
What if every time one saw red, the other saw purple; every time one saw 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.
Similar problems could be constructed for all sense modalities.
So, functionalism fails to capture the nature of our qualitative mental states.

The absent qualia argument goes one step further than the inverted qualia argument.
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 are a matter of degree, not of kind.
The basic picture of neurons transmitting information like electrons passing along a circuit board is apt.
The absent qualia argument is that there are functional equivalents of minds which lack properties that minds should have, in particular conscious properties.
Where identity theory was chauvinist, functionalism is too liberal, ascribing minds to too many things.

Ned Block, formulating the absent-qualia argument, considers 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.
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.

VII. Summary

The Chinese room argument, from Searle, shows that functionalism has a problem accounting for our intentional states.
The absent qualia argument, from Block, shows that functionalism has a problem accounting for our phenomenal states.
Only the dualist provided a satisfying first-person account.
But dualism has an apparently insuperable (and spooky) problem of interaction.
Functionalism is the most widely supported contemporary theory of mind.
Fodor observes that it deals with intentional content better than phenomenal content.

> Functionalism has fared much better with the intentional content of mental states. Indeed, it is here that the major achievements of recent cognitive science are found (Fodor 457a).

One live alternative to functionalism, called eliminative materialism, denigrates, like the behaviorist, our internal states.
We saw a little bit from the eliminative materialist in the quotes from Paul and Patricia Churchland about their brain states.
The functionalist can adopt some of the eliminative materialist's attitude about phenomenal states, denying that they are legitimate elements of any proper psychological theory.

We have one more, related topic in philosophy of mind to discuss.
We will look a little more at whether a physicalist theory of mind is what we really want.