Philosophy 340 W. Lycan
Spring, 2015

INTENTIONALITY, III

Searle vs. computationalism

"Strong AI" is the thesis that I can understand what someone says, or be in some other intentional state, *just* in virtue of performing formal computations over uninterpreted "symbols," really just syntactic shapes. Searle attacks Strong AI by offering his Chinese Room counterexample (very like Block's population-of-China alleged counterexample to Functionalism), in which there is formal computation over uninterpreted squiggles, but, Searle says, obviously no understanding or other intentionality.

Objections considered in class:

1. The "systems reply": That no component of the Chinese Room setup has understanding or (relevant Chinese) intentionality does not show that the system as a whole doesn't have it. The system is, after all, running the program of a real Chinese understander. We don't want to commit the Fallacy of Composition.

Searle would make two replies. In reverse order: First, let the Searle-homunculus in the example "internalize" all the other elements of the system, so that he becomes the whole system. He still understands no Chinese. Second, without committing the Fallacy, it seems nearly as obvious that the whole Chinese Room doesn't understand Chinese as that Searle qua homunculus doesn't. "It is not easy for me to imagine how someone who was not in the grip of an ideology would find the idea at all plausible."

2. The "robot reply": The trouble is that the Chinese room does not interact with its environment; it's completely passive, and only outputs Chinese characters in response to other Chinese characters.

Searle: OK, so add wheels and robot arms. So what? That won't make it understand Chinese either. We could go on and add video cameras and audio; but that will only give the Searle homunculus sights and sounds of the environment outside; it would have no tendency to induce understanding of Chinese.

3. The "many mansions reply": Searle was mired in 1970s technology. Machines and programming have massively improved, and there are new and stunning architectures, such as Connectionism, "dynamical systems," etc.

Searle: There's a dilemma. Either the wondrous new method in question is just another albeit souped-up species of formal computation over uninterpreted

symbols, or it works by adding something to that. If the former, it doesn't matter how new or wondrous the method is; the Chinese Room argument goes through as before. If the latter, well and good; Searle's target is only Strong AI. Perhaps then we'll know how a machine really could understand a story written in Chinese.

Tentative diagnosis: The key word is "*uninterpreted*" in the phrase "uninterpreted symbols." If the symbols are really *symbols*, they have meanings; it's just that the machine can't see those. The reason the machine can't see the meanings but only the shapes is that the meanings are relations between the shapes and things in the world; "green" refers to green things, "dog" means dogs, etc.

So, we can concede that Searle is right and Strong AI is false. What's needed for understanding is being properly connected to the external world. Jerry Fodor, as before, agrees that a "psychosemantics" is needed for a theory of human minds; what we need on Searle's behalf is a psychosemantics for the Chinese Room.

A typical psychosemantics for humans features some naturalistic relation between brain states and kinds of thing in the world—say, the *typical cause* relation. Take whatever will prove to be the correct psychosemantics; then just apply it to the Chinese Room. If it does apply, it shows that the system does after all have intentional states. If it doesn't, that would show why Searle is right and the system can't have intentional states. Either way, there's no issue left.

Searle rejoins: Egg foo yung! Connecting a squiggle inside the room causally or historically with a bowl of egg foo yung outside would not help, because it would not enable the Searle-homunculus to interpret the squiggle—he can't see outside the room. At the least, he would need to be *aware of* the causal relation.

Fodor would say that *the Searle-homunculus* doesn't have to know the squiggle's meaning, any more than one of your neurons needs to know what a brain state is representing; he's only a functionary. What matters is whether the whole system is connected to egg foo yung in the right way, and as before, either it is or it isn't. But Searle will rejoin that if the whole Chinese Room didn't understand Chinese or have other intentional states before the causal-historical chain was added, the addition of the chain does nothing to help. Stalemate.

(But remember, Fodor and Searle agree that Strong AI is false. Their disagreement is over Strong-AI-plus-naturalistic-psychosemantics.)

Notice the relevance for theories of mind: ("Psycho-")functionalism, the reigning materialist theory, seems to entail Strong AI. So if Strong AI is false as Fodor and Searle are now agreeing, so is Functionalism.

Granting that Strong AI is false, Fodor now admits that Functionalism is not true *of intentional content*. What makes the belief that my aunt has a new dog

a *belief* is as always its functional role. But what makes it the belief *that my aunt has a new dog* is its psychosemantic relation to the external world.

Methodological solipsism

Several different computers could be printing out the same numbers in response to the same keystrokes, yet computing different things. Putnam, Fodor and Stich argue that much the same is true of human beings, even though our intentionality is original rather than derived: Surprisingly, ordinary propositional attitude contents do not seem to be determined by the states of their subjects' nervous systems, not even by the total state of their subjects' entire bodies. Fodor's "Yon" (= Putnam's "Twin Earth") and indexical examples ("I am overpaid") and beliefs involving proper names ("George is underpaid") and "New Earth" are widely taken to show that, surprising as it may seem, two human beings could be molecule-for-molecule alike and still differ in their beliefs and desires, depending on various factors in their spatial and historical environments. This is called **externalism** about intentional contents.

Thus we can distinguish between "narrow" properties, those that are determined by a subject's intrinsic physical composition, and "wide" properties, those that are not so determined, and representational contents are wide. So it seems an adequate psychosemantics cannot limit its resources to narrow properties such as internal functional or computational roles; it must specify some scientifically accessible relations between brain and environment.

Fodor and some other theorists continue to maintain that there are narrow contents in addition to, and perhaps underlying, the wide ones. For example, although WGL believes that water covers 67% of the earth and Twin WGL believes something different, that Twin-water (XYZ) covers 67% of the (Twin) earth, each believes that *the familiar clear odorless tasteless liquid covers 67% of the planet he inhabits*.

(Quick quiz: That last bit was supposed to be an example of a narrow content that WGL and Twin WGL would share. Is it really an example of that?)

You may or may not be convinced that the Putnam-Fodor-Stich externalist thesis is right. But *if* it is right, it is highly consequential. Here are some morals that people have drawn from it.

Methodological solipsism for scientific psychology. Psychologists are supposed to be studying the causes of behavior. But molecular duplicates behave in the same way, as a result of the same causes. That my belief is about water (H_2O) while Twin WGL's belief is about a different stuff (XYZ) is causally irrelevant; we behave the same, because the causes of our behavior are in the head.

Accordingly, psychology should appeal only to what's in the head, i.e., to narrow properties of subjects and not their wide properties.

(In his comments on Fodor, Stich warns that there may be a lot fewer narrow contents than Fodor thinks, indeed very few for the psychologists to work with at all. What Stich really believes, though he doesn't come right out and say it here, is that psychology should get entirely out of the content business.)

Functionalism as we have been understanding it is untenable. Till now we have conceived functional roles narrowly; you and your Twin would have exactly the same functions being realized in you from moment to moment. Yet you believe and want different things; hence, believing and wanting are not functional states.

The Functionalist has two possible moves here. One is to restrict the thesis: Function is what makes a belief a belief, but some other account must be given of its intentional content. (We need a psychosemantics in any case.) The other move is to fashion a Wide Functionalism, and define functions themselves in terms of environmental entities rather than abstractly characterized body parts.

We think of propositional attitudes as having their characteristic effects in virtue of their contents: people behave as they do because of what they believe, what they want, what they plan, etc. But if those contents are wide rather than narrow, *that commonsense way of thinking is wrong*. You and your Twin will always behave in just the same way, despite having different beliefs and desires. Behavior is caused by what's in the head, not by the relations your head contents bear to things in the environment.

Luca had an excellent answer to this. In the spirit of Wide Functionalism, he pointed out that *behavior* can be characterized widely too (George bangs on the Dean's door, while Twin George bangs on *his* (Twin) Dean's door—different doors, you see, so different behavior despite sameness of bodily motion).

There is a problem about our knowledge of our own intentional states. We think we have secure first-person knowledge of what we believe. Our self-knowledge is not infallible, but it is privileged; you have a kind of direct access to your belief contents, in that in order to know what they are you don't have to examine any part of the external world. But if those contents are constituted in part by things and stuffs in your external environment, how is that possible??

<u>Psychosemantics</u>

Dretske bases his nascent psychosemantics on the notion of "information" that figures in the Shannon-Weaver "Mathematical Theory of Communication." A state of a device carries the "information" that P iff by law of nature, the device could not be in that state unless P. Of course, this sort of relationship is ubiquitous

in the universe, because there are lots of laws of nature and scads of nomic relationships. So Dretske's task is to say what must be added to "information" to get a genuine *cognitive* intentional content.

He focuses on the *most specific* "information" carried by a state. His second organism (p. 360) can know that it is touching acid without knowing that it is touching HCl because it has a repertoire of two different states, one of which carries the information that it is touching HCl (but nothing more specific) while the other carries only the information that it is touching acid. He calls this a difference in two ways in which the creature can "code" the information that it is touching acid.

That suggestion does not help much. Two thermostats or galvanometers could have such pairs of states. It follows from what Dretske says here that they would be genuine cognizers.

But this is only the beginning of a long tale of woe. The next problem is that which Dretske confesses on p. 361: How is he to accommodate *false* intentional contents, such as those of false beliefs? (If we were to say that a belief that P is even in part a state that carries the information that P, then the subject by definition could not be in the state unless it were true that P.) Dretske gestures toward the organism's learning history, but that idea never panned out.

In later work, Dretske injected an element of teleology: A state carries the content that P iff it *has the function of* carrying the information that P. That explains how a belief can be false; the belief is supposed to carry the information that P, but fails to. (Examples: A malfunctioning gas gauge, a watch that's fast.) That's helpful, but again is only a baby step. Objection (Johnathon and Luca): The new teleologized theory entails that every time I have a false belief, I'm broken or at least *malfunctioning*. But some of my false beliefs are not the result of malfunction. Sometimes I am in perfect working order and hold a belief that is perfectly well justified (indeed, I'd be malfunctioning if I *didn't* hold it), yet the belief is false owing to a fluke.

A worse objection: How does the gas-gauge, galvanometer model apply to the human brain? Are we to think that various individual brain states have as their *biofunctions* to carry information about the speed of light, about the Australian bushranger Ned Kelly, about God and my sister? That is a weird suggestion. Compare: The biofunction of teeth is to pulverize food so that it can be swallowed. If you like, the biofunction of eyes is to extract information from the immediate environment and to output a representation of what sorts of physical objects lie before one. But how could it be a biofunction of a state deep inside cerebral cortex to indicate something about the speed of light or Ned Kelly?

A further and much more serious obstacle to psychosemantics is that the objects of thought need not be in the environment at all. They may be abstract;

one can think about a number, or about an abstruse theological property. And as always they may be entirely unreal. (N.b., the same things are true of representations posited by cognitive psychology.) An adequate psychosemantics must deal just as thoroughly with arithmetical beliefs, and beliefs in quantum field theory, and Arthur's illiterate belief that the number of the Fates was six, and Hegel's belief that the Absolute is in a constant process of self-realization. I have no idea how present-day psychosemantics could be extended to deliver contents like those. (Though Mark made a pretty good suggestion: that perhaps the abstract belief contents are somehow composed of smaller informational contents that are less far removed from the perceptual.)

Finally: Dretske and other psychosemanticists focus almost entirely on *belief*. (I say "almost" because in his 1988 book Dretske does also address desire.) Of course belief is information-carrying in some sense; it represents the world as being a certain way, even if we don't agree with Fodor that it is a kind of internal, computational representation, and it aims at correctness. But what about other propositional attitudes whose function is not to be *correct* representations? S wishes that P; S wonders whether P; S hopes that P; S fears that P. It is not the function of any such state to carry the information that P. So here is another dimension along which Dretske's account will have to be extrapolated, and I don't see bright prospects for that project.