6. Behaviorism

Martín Abreu Zavaleta

June 4, 2014

1 More on Turing's test

Recall the three different attitudes one could have towards Turing's test:

1. Completely deny that passing the Turing test guarantees being able to think or being intelligent. At best, the Turing test is good for *simulating* thought and intelligence, not for *genuinely* having thoughts and intelligence.

2. Think that passing the Turing test gives us good reason to believe that a computer can think or is intelligent, but doesn't guarantee being intelligent or having thoughts.

3. Think that passing the Turing test *suffices* for being intelligent or having thoughts. According to this response, being intelligent *is just* to behave in certain ways. In particular, being intelligent *is just* to behave in the ways that something would need to behave in so that it passes Turing's test.

We already examined some arguments that people with attitude 1 may have. How about the other two attitudes?

Let's focus on 2 first. Suppose that the fact that computer C passed Turing's test didn't even give us reason to think that C is intelligent or capable of thought. Perhaps, following attitude 1, we think that it only shows that C can simulate intelligence. Now think about the kind of evidence that you have for thinking that other *people*, apart from yourself, are intelligent. Is it substantially different from the kind of evidence that passing Turing's test can give us?

When we attribute intelligence and thoughts to other people, all the evidence we have for this consists in the way they behave. We think that other people are intelligent because of the way in which they behave: they speak intelligibly, respond to our answers in coherent ways, and so on. So why should people's intelligent-like behavior give us good evidence that they *genuinely* are intelligent, but reject this for computers? There doesn't seem to be any principled reason.

Some people, those endorsing attitude 3, take this line of reasoning further. They claim that passing the Turing test would not only give us good evidence for the passer's intelligence. It would *guarantee* or *suffice* for the passer's intelligence. This is so, they claim, because being intelligent, having thoughts, and so on *is just* to behave in such and such ways. In particular, being intelligent, having thoughts, and so on, is just to behave in the ways in which something that passes the Turing test would have to behave.

People endorsing attitude 3 are usually called *behaviorists*. Sometimes they are called *logical behaviorists* to distinguish them from the defenders of behaviorist theories in psychology.

2 Behaviorism

Let's start with some questions. On certain conception of the mind, we have a special kind of access to the contents of our own minds, but not to the minds of others. For instance, you can know *first-hand* that you are in pain, or that you are having a certain thought. Others can't know this about you in the same way. In order for them to know that you are in pain or that you are having a certain thought, they need to look at what you do. If you say 'ouch', then this is good evidence that you are in pain, and if you say 'I think that roses are red', this is good evidence that you are indeed thinking that roses are red.

But what if when I say that I'm in pain, what I'm really experiencing is something more like a tickle, or even pleasure? On the view of the mind described above, it seems possible that I act exactly the same way I do, yet experience tickles when I claim to experience pains, and the other way around. After all, I can't be sure that my use of the word 'pain' is wrong, since the only sensations that I can know are my own!

These kinds of considerations worried some people in the past. They thought that the idea that this was possible attempted against the commonsensical idea that communication is in principle possible. We are not going to examine this more thoroughly, but we can summarize this in the form of a perfectly sensible worry: *if our behavior is to count as evidence of our mental states, then there must be some relation between them*—that is, there must be some relation between behavior and mental states, that explains why the former gives us good evidence for the latter.

Behaviorists think that behavior is good evidence of mentality (and of more particular mental states) because *having a mind is just a matter of exhibiting or having a propensity, capacity or disposition to exhibit appropriate patterns of behavior*(Cf. Kim, p. 62). So what is behavior?

It's hard to give a definition, but some examples should help: *physiological reactions and responses*, together with *bodily movements*, count as behavior for our purposes. For instance, an increase in pulse rate, raise in blood pressure, perspiration, raising a hand, walking, running, uttering noises, and so on. The crucial feature of these kinds of movements is that we don't need any psychological vocabulary to describe them. Contrast them with events that require some psychological component, like greeting your friends, calculating, or thinking about what you'll do when the class is over.

The crucial feature of behavior is that it's *public*: everyone has equally good access to someone's behavior, including the person whose behavior we are studying. Contrast with things like sensations or beliefs: the experiencer of these sensations or beliefs is supposed to have some special access to them, but she has no special access to her heart rate, or to the measure of her blood pressure, and even if there is something it feels like to walk, as long as she is walking, anyone who observes her movements is in as good a position to know that she is walking.

You can start seeing how, if behaviorism is true, it can explain why behavior gives us evidence of mentality. Since behaviorists reduce mentality to behavior, seeing that someone exhibits a particular behavior is, in a way, just like seeing that she exhibits mentality.

But notice that we are not always behaving in intelligent ways. Sometimes we sleep, and sometimes we just don't exhibit behavior that signals any complex mental states: for instance, when we're eating or yawning. Should we say that we are not intelligent when in those cases? That would be weird! This is why behaviorists appeal to the notion of a *disposition*. Let's define behaviorism in terms of dispositions, and then define the latter:

Behaviorism (improved): All there is to having some mental state is being *disposed* to behave in

certain ways in response to certain kinds of stimuli.

Now we'll say something about dispositions.

Dispositions

Consider properties like being fragile, being crushable, or being soluble. We say that something is fragile if it is easily broken, and soluble if it can be dissolved in some solvent. But of course, when we say that a glass is fragile, it doesn't *have to be* already broken, it just has to have a certain propensity to break. In order for something to be soluble, it doesn't have to actually been dissolved, it just has to be such that it dissolves if a certain condition obtains.

One way of refining the notion of a particular disposition is by means of *counterfactuals*, that is, sentences like the following:

1. If I had eaten enough at lunch, I wouldn't be hungry now.

2. If I had unplugged my fan, it would have stopped running.

The particular counterfactuals above are true, not not all counterfactuals are. For instance, here is a false counterfactual: if I had woken up earlier today, the Martians would have invaded us.

Now let's go back to dispositions. One way of defining what it is to be fragile is by appealing to a counterfactual like this one: something is fragile if and only if it *would* break if struck in certain ways; something is crushable if it *would* be crushed, had we applied a certain force to it; something is soluble if it *would* dissolve if it were placed in water.

Notice that in order for a glass to be fragile, it doesn't have to *actually* break. For a sugar cube to be soluble in water, it doesn't have to be *actually* dissolve. A glass could be fragile if it never breaks. For instance, I may make a glass so fragile that I worry it would break if even the slightest force struck it, so I put it in a safe box, and it never breaks in its whole history. Perhaps at some time it is disintegrated, but it never breaks. That doesn't make the glass any less fragile.

When people talk about dispositions, they often distinguish between manifestations and categorical bases:

Manifestation: An event of a glass being broken, or of salt being dissolved, are manifestations of their fragility and solubility, respectively. Some people call *manifestation conditions* to the conditions that have to obtain in order for a disposition to be manifested—e.g. for a glass to break.

Categorical basis: The categorical basis of a disposition is the set of properties in virtue of which something has the disposition. For instance, salt may be soluble because of its molecular structure. Even if different things have the same disposition, the categorical basis need not be the same for all of them. For instance a glass and a house of cards may both be fragile, but their fragility may be based on different categorical bases.

Now that you know what dispositions are, you are in a better position to understand the formulation of behaviorism above.

Very few people these days accept behaviorism, and with good reason. Several philosophers have criticized on these kinds of grounds:

(i) We all think that mental states have *causal powers*: feeling pain may cause us to move our hands away from the fire, and beliefs and desires may cause us to act in certain ways. It's not clear how behaviorism captures this intuition.

(ii) Some philosophers have offered the following kind of thought experiment: suppose that someone makes a doll that she can control remotely. She makes the doll behave in exactly the same way a human would. Does this mean that the doll has genuine thoughts or mentality? Most of us wouldn't think so.

(iii) Can we really reduce every possible mental state to a set of dispositions? Try defining beliefs merely in terms of dispositions to action!

We may try to define what a belief is solely in terms of dispositions. For instance, let's say that to believe that p is to have the disposition to utter the sentence 'p' when asked whether p. **Questions:** Do you think this is a correct definition of belief? why, or why not? Can you think of cases in which the definition would fail?

Further problems come from the fact that different kinds of mental states seem to stand in complex relations with each other, and this makes it very hard to see how a behavioral definition of the mental states could be given. For instance, consider the following principle: if a person desires that p and believes that doing A is an optimal way to secure p, she will do A. **Question:** What do you think about this principle? Is it true? Can you think of any situations that would falsify it?[1]

[1] One way of arguing in philosophy is by means of *counterexamples*. A counterexample is a case that falsifies a general claim. For instance, if I say that all swans are white, any swan that is not white would be a counterexample to my claim. If there are counterexamples to a general claim, that means that the general claim is false.