Is Consciousness Nothing More Than Input/Output?

Let us start by considering a story:

Raymond Smullyan, "An Unfortunate Dualist" (1980)

> Once upon a time there was a dualist. He believed that mind and matter are separate substances. Just how they interacted he did not pretend to know—this was one of the "mysteries" of life. But he was sure they were quite separate substances.
>
> This dualist, unfortunately, led an unbearably painful life—not because of his philosophical beliefs, but for quite different reasons. … He longed for nothing more than to die. But he was deterred from suicide by such reasons as … he did not want to hurt other people by his death. … So our poor dualist was quite desperate.
>
> Then came the discovery of the miracle drug! Its effect on the taker was to annihilate the soul or mind entirely but to leave the body functioning exactly as before. Absolutely no observable change came over the taker; the body continued to act just as if it still had a soul. Not the closest friend or observer could possibly know that the taker had taken the drug, unless the taker informed him. … [O]ur dualist was, of course, delighted! Now he could annihilate himself (his soul, that is) in a way not subject to any of the foregoing objections. And so, for the first time in years, he went to bed with a light heart, saying: "Tomorrow morning I will go down to the drugstore and get the drug. My days of suffering are over at last!" With these thoughts, he fell peacefully asleep.
>
> Now at this point a curious thing happened. A friend of the dualist who knew about this drug, and who knew of the sufferings of the dualist, decided to put him out of his misery. So in the middle of the night, while the dualist was fast asleep, the friend quietly stole into the house and injected the drug into his veins. The next morning the body of the dualist awoke—without any soul indeed—and the first thing it did was to go to the drugstore to get the drug. He took it home and, before taking it, said, "Now I shall be released." So he took it and then waited the time interval in which it was supposed to work. At the end of the interval he angrily exclaimed: "Damn it, this stuff hasn't helped at all! I still obviously have a soul and am suffering as much as ever!"

What is the moral of this story?

At the end, Smullyan hints at the moral when he asks us, "Doesn't all this suggest that perhaps there might be something just a little wrong with dualism?"

If dualism were the correct view, then it should be in principle possible for a being to exhibit the external behavior of consciousness without actually BEING conscious (i.e., "philosophical zombies" would be metaphysically possible). Smullyan clearly thinks that this is absurd. [*Is it?*]

He seems to suggest that, if an individual *behaves* in a way that is *indistinguishable* from other humans, then this surely guarantees that they ARE a conscious being.

1. The Turing Test: In 1950, Alan Turing suggested something similar for artificial intelligence. He proposed that a machine would truly be able to **think** if they could pass a certain test. (A nice video about it can be found here.) During the test, a human being would ask the machine a series of questions via something like text message. If the human couldn't tell whether she was texting a human or a robot, then the A.I. passed the test—that is, based on its behavior, we would know that the machine could THINK.

Out of this line of thinking, the following view about consciousness was proposed:

> **Behaviorism:** If something *behaves* like a conscious being – e.g., by demonstrating the appropriate inputs and outputs – then it *is* conscious.
>
> On the strictest version, conscious states JUST ARE the behavioral states. For example, to be in pain *just is* to say "Ouch!" in response to certain stimuli.
>
> More plausible version: To behave like a conscious being is the sort of *evidence* that guarantees that a thing is conscious. (*We'll discuss this version.*)

Note: This view is quite *amenable* to (but does not strictly entail) a **physicalist theory of consciousness** – that is, the view that conscious experiences are NOTHING MORE than purely physical events. For, presumably, we could design a purely physical thing to behave like a conscious being. And, if that alone guarantees consciousness, then it follows that we could design a purely physical thing that IS conscious.

This view also seems to entail that **artificial intelligences could be conscious** – namely, if they ever become advanced enough to behave like conscious beings. That said, this view is also a rejection of the **biological view** of consciousness – i.e., the view that only organic, biological brains can be conscious.

With innovations like Chat GPT, the day when A.I. easily passes the Turing Test seems to be fast approaching (if it's not here already). So, is *A.I. consciousness* fast approaching?

2. Objection: The Chinese Room: Consider the following case, from John Searle (1971):

> **The Chinese Room** You do not speak a word of Chinese. Some scientists stick you in a room all by yourself and give you a giant book, filled with thousands of pages of strange symbols. The scientists tell you that they will be slipping pieces of paper under the door with more weird symbols on them. When you receive these slips of paper, you are to consult your book and find that string of symbols in the left column. You must then write down whatever is written in the *right* column, and slip that paper back under the door. (Videos here and here.)

It turns out that you're actually receiving, and writing down Chinese characters, and having a conversation with a fluent Chinese speaker outside, who is convinced that they are conversing with someone who understands Chinese. And yet, you have no idea what you are doing. In a sense, **you are *behaving* like you know Chinese, but you do not**.

Against Input-Output as a Test for Consciousness: But, then, it seems to follow that, if a machine *behaves* convincingly – for example, if you interact with a chatbot A.I. and become totally convinced that you are interacting with a conscious being – this does not actually indicate consciousness. In fact, if the machine behaves anything like you do in the Chinese Room case, then it is obviously NOT conscious! For, just as you are merely *imitating* the way in which a fluent Chinese speaker would behave, in that case a machine would also merely be *imitating* the behavior of a conscious being.

And in fact, A.I.'s sort of DO act like you do in the Chinese Room. It's a bit more complicated these days, but imagine instead that you were given a HUGE number of the strange symbols as data sets, AND you learned to assign probabilities for how appropriate certain symbols would be, given certain inputs, etc. etc. Even though modern A.I.'s are more complicated than the one-to-one correspondence of input-output that Searle imagines, it doesn't seem to affect his overall point: **A.I.'s will never be conscious, they will merely *simulate* consciousness**.

(*To illustrate, it may help to simply interact with some chatbots, e.g., Chat GPT, or Kuki (formerly Mitsuku), or Sophie. Note: You will need to ask Chat GPT to pretend as if it is a human being during your conversation. Then: Ask yourself whether the bot would become conscious so long as it merely became **more convincing**. What do you think?*)

An argument for this conclusion would look something like the following:

1. **Behaviorist Claim:** Whenever something *behaves* as if it is conscious (convincingly, after close examination), then that thing *is* conscious. (Entails: If a thing behaves as if it understands X, then it *does* understand X.)
2. The man in The Chinese Room behaves as if he understands Chinese.
3. The man in The Chinese Room does *not* understand Chinese.
4. Therefore, behaviorism is false.

Objection to P3: Perhaps the *system* as a whole DOES have understanding? We would not expect the PERSON in the room to have understanding, because she is only a COMPONENT of a system that understands, just as we would not expect a set of neurons to have understanding, because they are only a COMPONENT of a system that understands. (*Searle asks us to imagine that the man internalizes the system – the book and instructions – so that he IS the system. He STILL wouldn't understand Chinese.*)

Objection to P2: Perhaps Searle's example is too simple. Imagine that the person in the Chinese room gains access to the outside world via cameras and microphones, and then begins to connect the symbols with the things seen through these cameras. She even begins to form NEW, increasingly complex strings of symbols. In short, imagine that:

(1) she connects the abstract symbols with things in the tangible world,
(2) she begins to learn on her own, and
(3) the original output that she creates on her own becomes very complex.

Would you say that the person in the Chinese room was still merely IMITATING the Chinese language? Or would we conclude that she now has TRUE UNDERSTANDING? (*Searle says that even here, while the MAN would perhaps gain some understanding of Chinese in this case, this is only because he is already conscious. But, if a MACHINE was connecting the signals delivered from the video feed to language symbols, this STILL would not constitute understanding, because it would still merely be running a program.*)

3. The China Brain: Perhaps the Chinese Room case is too simple. But, surely every functionalist would agree that – if a system functioned exactly like a HUMAN, including the brain with all of its intricate inputs and outputs – surely THAT sort of system would DEFINITELY be conscious. This is a version of functionalism:

> **Functionalism:** Merely *behaving* like a conscious being doesn't guarantee consciousness. Rather, that behavior must be the product of a particular kind of system that *functions* in a specific kind of way – and we know that one of those ways is the way that the human brain functions (though there might be other ways that a system could function that are sufficient for consciousness).
>
> Importantly: If a system *functions* in the appropriate way, then it *is* conscious.

But, imagine this case from Ned Block (1978):

> **China Brain** The entire population of China (1.4 billion) are asked to engage in an experiment. The motions of a giant robot are dictated by a network of radio signals. Every single person is given a walkie-talkie and a series of instructions. The range of the walkie-talkies only extends to those people nearest to them. The instructions are things like, "When the person in front of you signals you with a beep, signal the person behind you with a beep," etc. (Or something.) The point of the experiment is to perfectly simulate the sorts of signals that NEURONS give to each other in your brain. So, the population of China perfectly copies the functionality of a human brain, and the robot body that they are controlling performs the actions that its "brain" tells it to.

[*Note: Scientists have since discovered that there are actually **86** billion neurons in the human brain. But, in any case, cats are clearly conscious, and they only have 700 million neurons—so, then, just imagine that the China Brain functions like two cats!*]

If functionalism is true, then the country of China would BE CONSCIOUS in this case. Block expects that you will find this absurd. We might construct a similar argument:

1. **Functionalist Claim:** Whenever a system *functions* like a conscious mind, then that system *is* conscious.
2. The population in the China Brain case *does* function like a conscious mind.
3. But, the population in the China Brain case (or, alternatively, the robot that they are controlling) is clearly *not* conscious.
4. Therefore, functionalism is false.

[*Note that Searle ALSO thinks that functionalism is false. For example, he presents a case where a man manipulates water pipes that function like the human brain, and argues that this system would not be conscious. But, Block's example is clearer than Searle's.*]

Searle would say that, in China Brain, the system is instantiated in *the wrong kind of stuff*! Ultimately, he seems to endorse a **biological view** of consciousness, stating that

> "Whatever else intentionality[1] is, it is a biological phenomenon, and it is as likely to be as causally dependent on the specific biochemistry of its origins as lactation, photosynthesis, or any other biological phenomena. No one would suppose that we could produce milk and sugar by running a computer simulation of the formal sequences in lactation and photosynthesis, but where the mind is concerned many people are willing to believe in such a miracle …"

Objection to P3: Perhaps a mind WOULD arise in the China Brain case. Perhaps this is not so strange. Consider the ways in which collections of bees or slime mold (also here) make seemingly intelligent decisions, for example. Is it so terribly counter-intuitive to think that a sort of "hive mind" arises in those cases? Some suggest that the individual bees act as individual neurons in such a way that the entire hive becomes one conscious being. Perhaps this same thing would occur in China?

You might object, "But, can't find the consciousness. This person isn't experiencing it, and this person isn't either. Etc. I've looked everywhere, and I don't see the conscious mind anywhere!!" But, that's like saying, "I've looked all over your brain. This neuron isn't conscious, and this neuron isn't conscious. So, apparently this human isn't conscious!"

[1] Roughly, the mind's ability to represent, refer to, or be directed toward things in the world. For example, to *perceive* a hand, or *desire* a pizza, *believe* a statement, or *understand* an argument.

If we examine a brain at the level of the neuron, it would not seem like there was a conscious mind there. Similarly, if we just look at one little bee, it will not SEEM like there is a hive mind. But, that is just because the hive mind is **not at** the level of the bee. (Just like the conscious human mind is **not at** the level of the neuron.)

Reply: Sure, hives might ACT like a single being, just as an artificial body connected to the China Brain might ACT like a human. But, would it BE CONSCIOUS? Is there some feeling or sensation that it is like TO BE a hive? Is there some qualitative feel that is what it is like TO BE the China Brain? That seems absurd. [*Do you agree?*]

[*Note: Searle is open to the possibility that we might be able to create an artificial or synthetic thing that is conscious. Only, the created being would (most likely) need to be a biological thing, with all of the same biological causal processes as biological beings.*]

Final Note: Like Behaviorism, Functionalism is quite *amenable* to physicalism, but it does not require that we endorse physicalism. Why not? Answer: On the version we're discussing, **functionalism is not a theory of what conscious experiences ARE**. For example, it might turn out that substance dualism is true, and that all of the things that FUNCTION like conscious beings are just the ones that have SOULS. Or perhaps property dualism is true, and all of the things that function like conscious beings turn out to have mental properties. (This is actually David Chalmers' view.) Though, with Smullyan, you may of course find it *surprising* if it turned out that nothing in the universe was able to even *function* like a conscious being without either some non-physical properties or substances entering the picture – realize that one *can* be a functionalist dualist. (And functionalist / property dualists are quite common.)