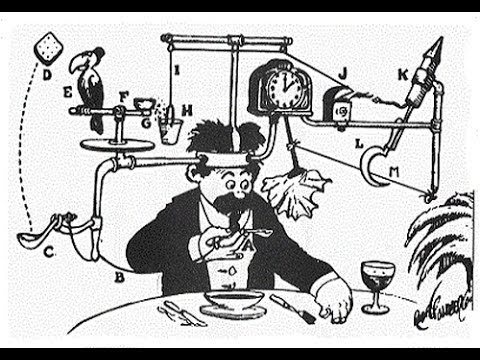
# Minds, Brains, and Programs: Functionalism and the Chinese Room

Three historically influential theories about the relationship of the mind/soul to the brain/body are as follows:

* **Dualism** holds that mind and the brain existed in entirely different realms, and that mind events (thoughts, pleasures, pains) are made to correspond with brain and body events (behaviors, the firing of neurons) by some outside force. The big problem? This theory can’t explain *how* the mind and the brain interact.
* **Behaviorism** holds the mental events are just behaviors. The big problem? This can’t account for “private” mental events at all.
* **The identity theory** holds that mental events are just brain events (in the form of specific neurons firing). The big problem? Critics have noted that this can’t account for **multiple realizability—**i.e., the fact that the same thought (“I am hungry”) might correspond to different neurons firing in different people. Moreover, if we discovered aliens, or built intelligent robots, they might not have neurons firing at all. However, many people have thought that at least some of these beings might have minds.

In this class, we’ll look at **functionalism,** which is probably the most popular theory of mind today (and has been for at least 40 years). Functionalism puts together key insights from each of these theories: dualism’s insistent on distinguishing the mental from the physical, behaviorism’s idea that sensory inputs produce behavioral outputs, and identity theory’s claim that that mental events all happen “in the brain” (at least for humans). Functionalism does this by comparing the relationship between the mind and the brain to the relationship between a computer program and the *particular* computer(s) the program happens to be running on. The key functionalist idea is that we should pay less attention to what minds are “made of,” and more attention to what minds *do.*

## Is the Mind to the Brain as the Program is to the Computer?

A **program** (or **algorithm**) is a set of purely formal rules for producing a specific output from a specific input. So, for example, the following program will print “Hello, [Your Name]!”

|  |  |
| --- | --- |
| Programming Language Command | What Happens |
| user\_name = input(“What is your name?”) | This line asks the person running the program to type in their name. This is a type of **input.** |
| output = “Hello” + user\_name + “!” | This command tells the computer to put together a new sentence using some rules. It provides a link between the input received in the previous step and the output that will be produced in the next step. (In real computer programs, this is usually much more complex.) |
| print output | This line prints “Hello [your name]” on the user’s computer screen. This produces **output.** |

This program can be run on all sorts of computer, regardless of the processor, operating system, etc. It doesn’t depend, for example, on whether the computer is a PC or a Mac, or even (in theory) whether the processor is made of silicon or some other substance. So, *the program is independent of the computer hardware.* Even more importantly, a computer doesn’t need to “know” English to process this program—the program can be run in a purely mechanical, rule-governed way.

**Functionalists** have argued that the human mind is, in essence, just a complex computer program, and that human brains are just the “hardware” that happens to run the program. For humans, the hardware in question happens to be neurons. But this didn’t need to be the case. Aliens or robots (or ghosts, angels, demons, etc.) might theoretically have the same sort of “minds” as humans (they might run the same programs), but they might run these programs on brains made of ectoplasm, or silicon, or whatever else you might imagine. Basically the idea is that the mind works like this:

|  |  |
| --- | --- |
| Mental “Program” | What Happens When Program is Run by th Braine |
| desire = survive + reproduce | The human brain forms and stores desiresof various sorts. Some of these desires are part of our biology, while others might depend on our social environment, upbringing, or whatever. |
| belief = input.sight + input.hearing + … | Humans use their five senses to take information about the world. This is stored in the brain as **beliefs** about the way the world is. |
| Behavior = intend(desire, belief) | Our **behavior** reflects our **intentions**, which are a function of our beliefs and desires. In humans this happens as the brain sends electrical signals along the nervous system. |

The basic idea is simple: A being has a mind if and only if it is capable of taking sensory input, processing it, and producing the “right” kind of output. Alan Turing developed the **Turing test** based on this idea. In the Turing Test, a person was given the chance to electronically “chat” with two other participants, one of which was human and one of which was a computer. Turing argued that, once computers were able to fool the human judges around 50% of the time, we would have to grant that they had minds. After all, they would have shown themselves to be just as good at conservation as humans, who we would all grant have minds.

## The Chinese Room Argument



Figure 1 From http://sophoslogos.wordpress.com/tag/chinese-room/.

John Searle wants to show that functionalism is false—that is, he wants to show that the mind is not merely a program run by the brain. He AGREES that the human mind can be modeled or described by a computer program. So, for example, Searle is perfectly happy with the functionalist idea that the human mind gets inputs, does something to them, and produces outputs. He also agrees with the functionalist claim that we might be able to build digital computers to run programs of very similar sorts, and that these programs might do a lot to help us understand and predict human behavior. However, as Searle notes, there are many processes can be modeled by computer programs, including the weather, the human digestive system, the flight of golf balls, etc. So, he doesn’t think we should be too impressed by the sort of argument given above (where we compared human minds to computer programs).

Searle’s DISAGREEMENT with the functionalist relates to their claim that that the mind just *is* a program, or that these sorts of analogies (between computer programs and minds) really help us understand anything about the relationship between mind and brain. Searle’s famous “Chinese Room Argument” is intended to show (in a much more concrete way) just what the functionalists gets wrong about the mind. In particular, Searle uses the thought experiment to argue that a machine running a computer program could perfectly capture the “inputs” and “outputs” *without thereby having a mind.* If this really is the case, then functionalism must be wrong, and human minds aren’t just computer programs. Here is the argument:

Imagine that you speak English well, but do not speak Chinese at all. You are hired to do a strange job, however. You stand in a room with two slots: one labeled INPUT and one labeled OUTPUT. You also have a book called PROGRAM. Chinese speakers pass in questions through the INPUT slot. You then look to see what the answer should be in the PROGRAM book, write down the answer (in Chinese characters), and pass through the OUTPUT slot. Throughout this process, you learn literally nothing about the Chinese language. However, to the Chinese speakers passing you the questions (and getting the answers) it seems as if you speak perfect Chinese. In fact, they think you do at least a good job answering questions as a native Chinese speaker would do in the same situation.

What exactly this thought experiment shows (or doesn’t show) has been a subject of considerable debate. With that in mind, here are some of the main things to remember:

**What does Searle thinks this example shows?** He thinks that the Chinese Room is a perfect example of a computer program that *appears* to have a mind. It takes questions in natural language and offers responses in natural language. So, according to a functionalist, the Chinese Room has a mind. However, Searle claims that it clearly does not. So, he concludes functionalism is false.

**What is the “systems response” to the Chinese Room Argument? Why does Searle think this doesn’t work?** Many functionalists give the “Systems Response.” It goes like this: Sure *you* can’t speak Chinese. However, the system of which you are a part (the room with two slots and the PROGRAM book) *can* speak Chinese. In this sense, the Chinese room DOES have a mind. Searle that this response fails. Why? Because he is convinced that reader will agree with him in saying that the Chinese Room does NOT “know” Chinese once they discover that the person in the room didn’t know any Chinese, and was simply following a PROGRAM book written entirely in English.

**What kind of machine *could* have a mind? Is Searle trying to defend Cartesian dualism?** Searle argues that it is much, much more difficult to create a mind than functionalists think. So, for example, he thinks making a robot that walks/talks wouldn’t (necessarily) have a mind, even if it ran on a computer program that “simulated” the action of the neurons in the human brain. That being said, Searle thinks it must be *possible* to create a “machine” with a mind, since we ourselves just are such machines. He just thinks that computer science isn’t anywhere close to having any idea what this might take. So, Searle isn’t a dualist (e.g., he’s not arguing for the existence of an immaterial soul); it’s just that he thinks the relationship between mind/brain is much more complex than functionalists are willing to grant. Making a mind, on Searle’s view, would require paying attention to the brain itself—what it is made of, how it works, etc. This would presumably require input from many people besides computer scientists or philosophers.

**What is the difference between Strong AI and Weak AI? What does Searle think of each?** According to **Strong Artificial Intelligence** human minds are just complex computer programs that run on the “hardware” of human brains. With sufficient skill and processing power, this means we could write a computer program for a silicon computer that also has a mind. Searle thinks the Chinese Room Arguments shows that this is false. By contrast, according to **Weak Artificial Intelligence**, humans minds can be *represented* by programs. With sufficient skill, we can make computer programs that can teach us about some aspects of human cognition. Searle thinks that this is fine.

## Review Questions

1. Carefully describe what exactly *functionalism* is, and why many people think it seems like an attractive theory of mind. Make sure to use your own words, and to give examples where appropriate. Then, explain how Searle’s “Chinese Room” argument raises problem for functionalism.
2. The Human Rights Commission has recently argued that Searle’s argument means that it would be immoral for us to try and develop “fully autonomous weaponry” (for example, drones that could independently choose targets and decide whether to fire on them). Why? They have argued that only a computer system with a true “mind” (that is, a system with Strong AI) would be capable of having empathy, and empathy is needed to make these sort of decisions. Do you agree with this argument? If you were in charge of weapon development for the military, would this sort of argument convince that you should stop research on these sorts of weapons?