Week 13 Notes
Philo 101 Online | Hunter College

Daniel W. Harris

1 How Could Materialism Be True?

Last week, we focused on the debate between Cartesian Dualism (the idea that minds and bodies are fundamentally different kinds of things that somehow interact) and Materialism (the idea that everything, including minds, is physical). This debate is ongoing within contemporary philosophy, but materialism is certainly the more popular theory. As we saw last week, there are still some significant challenges to materialism being debated now. Nonetheless, this week our task will be to assume that materialism is true, and ask how this could be the case, and what would follow if it is.

In particular: if our minds are physical entities, how do they accomplish all of what they do? How could all of our thoughts, emotions, sensations, and conscious experiences be the sorts of things that ultimately boil down to physical processes?

For the last several decades, the most widespread answer to this question has been the *Computational Theory of Mind* (which I will abbreviate as 'CTM'). The central idea of this theory is that minds/brains are, quite literally, computers. Of course, this idea requires that we define 'computer' somewhat broadly, as a device for processing many different kinds of information. If that's what a computer is, then CTM says that a mind/brain is a computer, and that it works on the same basic principles on which other computers work.

You might be asking the following question: How could the mind *literally* be a computer? Computers are made out of circuit boards and storage media, and minds/brains are made out of living cells. Aren't those just different kinds of things?

To answer this question, we have to think of computers not in terms of what they're made of, but in terms of what they *do*. And this means that we should think of computers at a higher level of abstraction than the question presupposes. To do this, it can help to recognize that even two computers can store and process information in ways that are physically very different.

Take, for example, the text file I am writing at this moment. I am composing it on a Mac laptop that stores it on a solid-state drive (SSD). But it is also being simultaneously backed up by Dropbox, which means that it is being sent over the internet to one of their data centers, where the file is being stored on a gigantic server's hard disk drive (HDD). SSDs and HDDs store information in physically different ways. A HDD stores information on a series of thin magnetic disks that spin very quickly (usually around 5400–7200 rotations per minute) so that the information stored on them can be read and written by a mechanical arm. By contrast, an SSD has no moving parts, and instead stores information in tiny integrated circuits. So, the very same information (this text file) is being stored in two entirely different physical media. What makes it the same file in both case? The answer is that the same *information* is stored in both media. More concretely: computers store information in binary, as a series of 1's and 0's. HDDs represent these 1's and 0's by changing the states of magnetic disks, whereas SSDs represent them by changing the states of their integrated circuits. But, in the case of the present file, these two devices are in states that match, in the sense that each is storing the same string of 1's and 0's.

The point of the foregoing paragraph was to illustrate, in some detail, that two different computers could function in the same way—that is, they could store and process just the same information—despite being made of very different kinds of components. Proponents of CTM think that this is a very important point, because it shows that our minds could be yet another physical medium in which information is stored and processed. Your beliefs, desires, pains, and emotions could all be elements of a very complex piece of software that is currently running on your brain, which stores all of this information in the states of your nervous system, rather than in the states of magnetic disks or integrated circuits.

On this view, the job of neuroscientists, psychologists, linguists, and other cognitive scientists, is to reverse-engineer the hardware and software that makes the human mind/brain work. This is a little bit like what a competitor of Google might do in order to develop a search algorithm that works like their's does. This competitor doesn't have direct access to Google's code, and so they have to infer how the code works on the basis of observing what the algorithm does in response to various inputs. Similarly, cognitive scientists don't have the ability to directly read the code that is running on the human brain, and so they have to infer how it works by doing experiments, collecting lots of observations, and theorizing about the mechanisms that are responsible for what they observe.

2 Artificial Intelligence

Although not all materialists believe in CTM, many think that it gives us our best chance of understanding how the human mind works. In practice, the way that many cognitive scientists investigate the human mind is by developing computational models of its different components. For example: computational linguists develop algorithms that mimic the ways in which humans learn and use natural languages like English, and vision scientists develop algorithms that mimic the ways in which humans process information about the light hitting their retinas in order to construct 3D representations of their environment. Although these projects are a long way off from complete success, they seem to be making remarkable progress toward reverse-engineering some parts of the human mind.

Projects like these also raise several kinds of question about *artificial intelligence*: If we get good enough at teaching a computer to mimic our thought processes, would we have developed an intelligent being, or just one that is good at mimicking our intelligence? Would we have created a machine with its own mind? What if this computer could think *better* than us in some ways?

These questions come into better focus if we distinguish three different things that people sometimes mean by the term 'artificial intelligence'. (This taxonomy is discussed in greater detail by Kukla and Walmsley in one of this week's required readings.)

> APPLIED AI
> The use of computers to do things that previously required human intelligence.
>
> WEAK AI
> The use of computers to precisely simulate human thought processes.
>
> STRONG AI
> The thesis that if we could develop a good enough weak AI, we would literally have created a mind.

This week's two optional readings are about the current state of applied AI. You have probably heard many shocking and perhaps worrying claims on this topic: Artificial intelligence has begun to advance at an increasing rate! It will soon do more and more of the jobs currently done by humans! It could begin to reproduce and evolve, and may eventually enslave us! Our reading by Lewis-Kraus emphasizes some of the impressive advances that applied AI has recently made. The reading by Gary Marcus points out that what we have now is still very limited in many ways, on the other hand.

The distinction between applied and weak AI is a very important one to keep in mind. To see the difference, notice that some of what we now call artificial intelligence is much better at certain tasks than humans in certain ways, and much worse in other ways. For example: Google Translate can produce somewhat useful translations between dozens of different languages—something that no human can do—but it also cannot understand certain uncommon phrases, such as very large numerals, that even small children can understand. So, although Google Translate is a very useful—and maybe intelligent?—piece of technology, it is not a good model for understanding how humans use language. In other words: it is a paradigmatic example of applied AI, but a very poor example of weak AI. This is because the purposes of applied and weak AI are different. Because the goal of weak AI is to model human intelligence, a successful weak AI would have to precisely mimic both the strengths and weaknesses of human minds. The optional reading by Gary Marcus points out many of the ways in which our current AIs are bad at things that humans are good at, but it's also worth thinking about the ways in which our current AIs are better at some tasks than us. Either of these differences gives us evidence that these systems aren't good models of how humans think.

Likewise, the distinction between weak AI and strong AI is important to keep in mind. Weak AI is a research project—the project of coming up with a computer program that can precisely simulate the activities of the human mind/brain. Strong AI is not another research project, but a claim about what we would accomplish by succeeding at the goal of weak AI. According to the strong AI thesis, a successful simulation of a mind would itself *be* a mind.

As Kukla and Walmsley explain, this is not always the case: a meteorologists's simulation of a weather pattern—however successful the simulation—is not itself a weather pattern. Why might we think that minds are different than weather patterns in this way? The answer, according to proponents of strong AI, is that the computational theory of mind is true. If so, then the minds are just very complicated pieces of software that run on our neural hardware. And if that is true, then a cognitive scientist who reproduces this software on some other hardware would also be recreating the kind of mind that they are modeling. Put more simply: if the mind/brain really is just a computer (i.e., if CTM is true), then a completely accurate computational simulation of a mind would itself *be* a mind.

3 The Turing Test and the Chinese Room

Two of our required readings for this week, by Alan Turing and John Searle, are about the question of whether the strong AI thesis is true. Turing thought so, and Searle thinks not.

Turing, who some consider to be the creator of computer science, is also among the first to propose a version of the strong AI thesis. In 'Computing Machinery and Intelligence', he proposes that the question of whether computers can think should be replaced by the question of whether a computer can perform as well as a human in what he calls 'the imitation game'.[1] If you're playing the game, your job is to figure out which of two individuals is male and which is female, just by asking them questions and reading their written responses. Whereas the man's job is to get you to guess correctly, the woman's job is to trick you into guessing wrongly. (It could just as easily be the other way around.) Turing proposes that if we replace one of the individuals with a computer, their performance in this game would be a good indicator of their intelligence.

These days, it is usually a simplification of this scenario that gets referred to as 'The Turing Test'. In this version, your job is to decide who is the human, and who is the computer, just by asking questions and reading written responses. The human's job is to get you to guess correctly, and the computer's job is to trick you into guessing wrongly. Turing's hypothesis is that any computer that could trick you around half of the time—thus making your guess no better than a coin flip—would be a genuinely intelligent machine.

Most of 'Computing Machinery and Intelligence' consists of responses to potential objections to this thesis. Part of your job this week is to understand these objections, and to thereby understand Turing's point of view as well as you can. Then you must do your best to understand and summarize an objection that Turing never anticipated—Searle's "Chinese Room Argument". Searle summarizes this argument in his essay, "Can Computers Think?" Your task is to summarize this argument, and then to imagine and explain how you think Turing would respond to

[1] *The Imitation Game* is also the name of a recent movie about Turing, in which he is played by Benedict Cumberbatch. Turing's life was fascinating, and the movie is pretty good!

it.