The Chinese Room Argument and Its Critics

In the last lecture we considered the Chinese room and began to discuss some of the objections that have been made to it. I want to begin this lecture by saying what I think is, at least part of its philosophical and scientific significance, and then we'll explore in more detail some of the objections, many objections that's been made to it and the way the argument has gone on, for now, fifteen years. It's important to distinguish, consciousness, our subjective states of awareness, from our understanding, or our ability to represent objects and states of affairs in out world. Philosophers have a technical term for this capacity of the mind to direct itself to objects, and states of affairs, it's called Intentionality And Intentionality is somewhat misleading, because it makes English speakers think that it's got some special connection with indenting, like most confusing philosophical words we borrowed it from the Germans. It's not a problem in German, but in English you gotta remember that Intentionality refers not only to intending, but: believing, hoping, fearing, desiring, wondering, whether, ah seeing, ah actually, ah actually carrying out an intention to do something, all of those mental states, the mental phenomena there are called Intentional, and the phenomena of Intentionality is supposed to be separable from the phenomena of consciousness, for the obvious reason that many of our intentional states are unconscious. Often we have beliefs, and hopes and fears, and desires of which we're not conscious. Now in the early days of cognitive science people wanted to separate the problem of Intentionality from the problem of consciousness, and a lot of people would have agreed, that computers, at least the ones we've built so far, aren't conscious, but they thought that they do have understanding, they do have Intentionality. Now part of the power of The Chinese Room Argument was it doesn't rest consciousness, it is not designed to show that computers aren't conscious—I think they're not—but that's a separate issue. The point is the power of it is that it shows they don't even have Intentionality. They don't have any form of understanding, not just in virtue of implementing the program. Now arguments have a logical structure, and I wanna make the logical structure of The Chinese Argument fully clear, because a lot of the criticism didn't really understand the structure, it's really a very simple structure.

The first premises of the argument, is the definition of a program. It follows from Turing's definition of a program that the program is defined in terms of the manipulation of formal symbols. I gave you the example of zeros and ones, but they needn't be zeros and ones, any system of formal symbols, where the machine can identify a symbol solely in virtue of its shape, or form, is adequate for a computer program. And I summarized that part of the definition of the program, using the linguist jargon, I say that programs are syntactical, that's by definition, the program is defined in terms of the operation on formal symbols, independent of the medium in which those symbols are realized. That's what I talked about when I said the symbols can be realized in silicon chips, or water pipes, or cogs in wheels, or pidgins pecking anything at all, provided that it carries out the steps in the program. But now there's another point we know from our own experience, namely, our minds have actual mental contents. When we think we do use symbols to think with. I tend to think in English, though eventually I got to point where I could think a little bit in French, but the point is the symbols that I used to think with, have a meaning, or a semantics, so minds have more than a syntax, they actually have a mental content, or a semantics. But now, what The Chinese Room Argument reminds us of, is step there in our argument, syntax by its self, just shuffling the Chinese symbols, doesn't so far, carry any meaning, there's so far no meaning attaching to symbols, that is to say, syntax by itself, is not the same as, nor is it sufficient for semantics. The symbols are one thing, the meaning is another.

But now from those three it follows, from one programs are syntactical, two minds have semantics, and three their not equivalent, the syntax by itself isn't enough for the semantics, it follows that programs by themselves are not sufficient for minds. That minds are not the same as programs, but that's just another way of saying that, Strong A.I. is false. That is from the fact that the program isn't by itself sufficient for, nor constitutive of a mind, we have shown that Strong A.I. is false, because Strong A.I. is simply defined as the view that minds are a species of program. That having the right program, with the right inputs and outputs is sufficient for having a mind.

Now as I've said, I thought that was a kind of obvious argument, and I didn't see any reason why it should meet with much dispute, but I have to tell you that it opened a Pandora's Box of debates, and those continue. There are a large number of, not very good books, and even worse PhD thesis that have been written on this topic, and in fairness to my critics I feel I have to go through a discussion of at least some of those, so you will understand the nature of the arguments against me, and I think that's not just that is a matter of establishing and argument, but there are genuine philosophical and scientific issues that surround this debate, and I want to illustrate some of those in the course of answering the objections.

I think what motivates the objections, at bottom, is a kind of residual Behaviorism. Its a kind of a feeling that if it walks like a duck, and it talks like a duck, and acts like a duck then it's gotta be a duck, and that's a mistake. That is, from the fact that a machine can talk as if it understood Chinese, and answer questions in Chinese, and behave as if it understood Chinese, it does not follow, that it understands Chinese, because understanding is something that goes on inside the mind, and it carries a mental content, and not just a formal syntax. However, I'm going to seriatim, as a series of steps go through some of the most famous arguments against the Chinese Room; so of the most famous arguments against the Chinese Room, some of the most famous answers.

The first one that I heard I thought was very revealing. One of the leading figures in the field, said to me; but look, suppose we program you in the Chinese room to answer the question, do you understand Chinese? So they give me this question in Chinese, do you understand Chinese, and I give back the answer—yeah, sure thing, I understand Chinese. Now what does that prove? I think it proves nothing! Let's go though it. I get in a symbol that looks like this, you're gonna have to forgive my Chinese writing, this is a dialect of Chinese you probably wont recognize, I get in a symbol that looks like that, I don't know that it means, it's a meaningless symbol, and I shuffle a bunch of other symbols unknown to me, that symbol means "do you understand Chinese," I don't know that, I look it up and I give back another symbol, in the same dialect, that looks like that, and unknown to me that symbols means, "why do you guy asks me these dumb questions, can't you see I understand Chinese, I've been answering questions in Chinese all week," a lot of semantic content packed into that symbol. OK, what does that show, nothing! That is the fact that I can behave systematically, as if I understa--stood Chinese, and can even answer the question, do you understand Chinese, in the affirmatively, in the affirmative, means nothing about my understanding of Chinese. Well let's, with that in mind, go through some of the other replies. I mentioned the systems reply, but another famous reply I call the Robot Reply, and here's how it goes.

Imagine, not a computer in a room by itself answering questions in Chinese, but imagine that we built a robot, and the robot could answer questions in Chinese, but it clanks around in the world, I mean it actually bangs up against things, and lifts things, and walks around—wouldn't the robot, just in virtue of the program, come to understand the words of Chinese, or any other language, because the robot is engaged in Causal Interaction with the rest of the world? I think the robot is no better off than the original machine in the room, why? Well, imagine, vary the example that I gave a little bit. I imagine myself in the room, now we imagine a robot, but imagine it's a great big robot and inside the robot's cranium is a room, and guess who's in that room? Me, I am in the room, and I get in symbols that come in from the robot's television cameras, it's has television cameras, and then it convert—it has transducers that converts the information from the television cameras into a bunch of meaningless, Chinese symbols which are meaningless to me, they come into the room and I process them, unknown to me I am getting symbols from the robots' sensory apparatus, and I'm giving out symbols to the robot's motor apparatus that enable it to clank its way though the world, but I'm on a special set of springs, I don't even detect the motion. I'm there in a locked room, I can see nothing, and I am processing these symbols. I am the robot's homunculus; I'm the little man inside the brain of the robot, but unlike the homunculus of classical philosophical theory, I don't understand anything, and if I don't understand anything, neither does the robot, because there's nothing in the robot that has a locus of understanding, but me, carrying out the steps, in the program.

Well, another reply which I thought was also revealing, and this gets closer to the nitty, re-get, gritty, was; look, it's true, our existing programs aren't sufficient for understanding, but that's because they don't simulate the right features of understanding, but suppose we had a really complicated program that actually simulated the behavior of the Chinese brain, we had a program that simulated the behavior of the brain, you could imagine doing it down to the last synapse, go to the neuron, or synapse level, whatever's your favorite level, you just simulate the behavior of the Chinese brain. Now then, the argument goes, if that program doesn't understand Chinese, then you'd have to say that native Chinese speakers don't understand Chinese, because what we're doing is having a computer program that simulates the behavior of the Chinese brain. Now I think that that argument is very revealing, because what it does is it combines the behaviorism that we saw earlier, with a kind of formalism that we've been seeing; that says the formal symbols are all that's necessary. That argument can't be any good and let me illustrate why. Suppose somebody said, look if you wanna build a machine that will digest pizza, just do a computer simulation of the digestive processes that go on in your stomach. But the computer doesn't digest pizza, what it does is produce model, or a picture of the digestion of pizza. If you've got this computer simulation running, and you rush out and buy pizza and stuff it into the computer, it isn't going to digest it, because digestion names an actual, causal, physical process, and what the computer does is formal model, or simulation, or description of that process; and I want to say, what's true of the computer simulation of digestion is true of the computer simulation of cognition. Even if you simulate it right down to the last neuron and synapse, you're simulating the wrong thing. That is to say, all that the computer does by definition, because programs are syntactical, is simulate the formal syntactical structure, but when we talk about the operation of the brain, we're talking about very specific causal processes. Very specific processes that actually cause states of consciousnesses, and you don't reproduce those states, by doing a formal simulation in terms of symbols. Now I would have thought that point was obvious, but a lot of people don't see it, so let me hammer it home with a couple of examples.

We don't know much about how the brain works, I mean we're making some progress, but we don't really know very much. But there is some things we do understand. We understand the effect of certain drugs on your nervous system. Now, not all of them, we don't know why alcohol makes you drunk, or aspirin cures your headache, but we know some of the things that cocaine, does. One of the disastrous things it does is that it impends the capacity of the synaptic receptors in the brain, to reabsorb a certain neurotransmitter, it's called norepinephrine, and the effect of this is that because the neurotransmitter isn't reabsorbed, it tends to stay in the synaptic cleft, and this has a dramatic effect on people's state of mind. OK, so you get the picture, we've got this neurotransmitter, it's squirted into the synaptic cleft, and it's, I-I-I was gonna say that it was sloshing around, but we're talk'in about, huh-huh, huh-huh-huh-a-a-a small number a molecules, I mean hardly a waves of this stuff, but anyway it's very small amounts of it. Now, here's the point, we could do a perfect computer simulation of that. That is to say, we get zeros and ones to stand for the different features of the process, by which cocaine effects the capacity of the brain to reabsorb neurotransmitters, but the computer simulation doesn't get a cocaine high, and if you have any doubts about that, think of it in terms of, of-a-a, actual models, I mean, let's do a computer simulation of, let's do a simulation where we use beer can and ping-pong balls, and let the beer cans represent the anatomical features of the brain, and the ping-pong balls can, different colored ping-pong balls represent cocaine, and norepinephrine. Now this system of beer cans and ping-pong balls doesn't feel anything because it doesn't have the right machinery. You can do a model, or a picture, of the operation of the brain at any level you like. Just as you can do a model of water molecules using colored ping-pong balls, or golf balls, but you can't swim, in a whole lot of ping-pong balls, because what you've got there is a model, and not the real thing. And why not, why don't you have the real thing? The answer is the one thing that's left out of the whole computational story-CAUSATION, the brain is a causal mechanism, the brain actually causes conscious states, and that's why a picture, of the processes by which the brain causes mental states, that's why a picture of the processes by which the brain causes conscious states, even though it might predict, or model, or simulate anything you'd like down to any level of accuracy you'd like, will leave out the essential element, if it does not capture the actual thing that brains have.

Now I'm gonna add this as a separate premises here, cause it's gonna be important, for the argument later. We're gonna say as a matter of how it actually work, how it actually works, you wanna say, brains cause minds, and that's just a short-hand for saying, neural-biological processes in the brain, as far as we know at the level of neurons and synapses, cause all of our conscious life. Keep that in mind, cause we're gonna come back to it, that's an important addition, to the structure of the argument.

OK, so far we've considered three objections then: The Systems Reply that we talked about the last time. The Robot Reply, and The Brain Simulator Reply.

Now, in fairness to my critics there have been a whole lot of others, and I can't talk about all of them, but I'll talk about the ones I think have been the most influential, and the most widespread.

One that' frequently made, and of course is quite correct, is that, in real life you couldn't build a program, that would allow me to pass the Turing test for understanding Chinese, if I was locked in a room. The truth is, you can't design a program that will enable a commercial computer to do that, we're no where near being able to do that, but I'm assuming, for the sake of argument, that in principle we could, in principle we could program a computer, so it can pass the Turing test for understanding a natural language. Now, I quite agree in real life, we, even if we had such a program, human beings couldn't carry it out in real time, it would take millions of years to go through all of the steps—but that's not the point. The reason we have these thought experiments, in science, and philosophy, is because there is lots of things, lots of experiments you can't perform in real life. I mean, Einstein's' famous experiments where he imagined we go to the nearest planet, in a rocket ship that goes ninety percent of the speed of light, that's a fantasy, we can't do that in real life, but it's a useful thought experiment. The reason I had the thought experiment, is not because I thought, we ought to build a room and lock me in it, and see if we could actually do this, I don't have the patience, or the life for that, but because I wanted to illustrate a deep-point. Even if we could, the syntax by itself is not going to be sufficient for the mental content, or the semantics.

Well, another objection, and this is an objection that I think is very, revealing, because it suggests to me that people don't really know what a computer is, and that objection goes as follows.

We're just talking about the existing state of technology, and many people say, yes, we agree, existing computers can't have Intentionality. They can't really be conscious, have intentional states, or understand Chinese, or anything else. But wait until next year. It's like, the old Brooklyn Dodgers, wait till next year, we're gonna have better computers next year, and they will actually have understanding. But, what's wrong with that, is that that argument, The Wait Till Next Year argument, suggests that we're talking about a particular state of technology, whereas the whole point of the example that I've presented you, is it rests on the definition of the computer, it has nothing to do with the state of computer technology, at any given point, it has to do with what a computer is, by definition. You see what technology gives us is, faster ways of implementing the syntax of the program, faster way of carrying out the computational steps. So the beauty of The Chinese Room Argument, is that it has nothing to do, with any state of technology, it's independent of any state of technology. And of course, you could always redefine the notion of computation, but computation is a well-defined notion, defined by Alan Turing in ways that I sketched earlier, and it that's the definition of computation, then we know as a matter of logic, that computation by itself, can't be sufficient for understanding. Because computation is defined purely formally, and that has nothing to do with any state of technology.

Alright, another objection, and I'm losing count here, a-what are we up to, about six or so, another objection to The Chinese Room Argument, was, well, a-and this objection proposed by Paul and Patricia Churchland, a-another objection they have recently put forward is, look, you might as well say a, that em, ah-the electromagnetic radiation can never be sufficient for light, cause if you go into a dark room, you can have machines that detect electromagnetic radiation, but don't detect light; cause what we call light is only a certain small fragment of the electromagnetic range. Now they say, why is it any different with syntax, maybe you're just in the luminous room, that the Chinese room is like the luminous room, and the guy's in there in the dark and he thinks, well, light can't be any electromagnetic radiation cause we got electromagnetic radiation and no light. Now that's a bad analogy for a very simple reason. What we're talking about when we talk about electromagnetic radiation, is an actual causal property, of a certain type of physical phenomenon to affect our sensory apparatus. What we call light is the effect of electromagnetic radiation on our sensory apparatus and on other optical devices, but this is the point, syntax as such has no causal powers. Syntax consists of purely formal symbolic devices; zeros and ones, or some other symbolic device, and the only causal power, is the causal power of the physical medium. The only causal power that the computer has is to go to the next stage of the program, when the machinery is actually running, but the syntax by itself, unlike electromagnetic radiation, has no causal powers at all. It's only the implementation which has causal powers.

See there's an odd feature that's lost in much of these debates, and I wanna emphasize it, and that's this. The brain is a physical organ, the brain is a kind of physical machine, and it has physical machine processes. Neuron firings at synapses are a kind of mechanical process. They are a machine process. Ironically, in their urge to be materialistic, the computationalists are not nearly materialistic enough, because computation does not name a physical process, computation names an abstract formal—mathematical—process, that we have found ways to implement on actual machines. So the kind of computer you buy in a store, is a kind a machine, but it's an electronic circuit, but computation names an abstract mathematical process that you can implement, or realize, on the machine, but computation does not name a machine process in itself. Anything can be computational. See, this is a deep point and it has to do with the nature of computation. Computation exists relative to an interpreter who assigns a computational interpretation. Watch, I'll show you a very simple computer.

This object is a very simple computer. It has a one step program. The program says stay there and don't fall off. OK, that's the program, and it happens to be carrying out the program, I can easily upset the program because it's a rather Mickey Mouse computer. But now this is an important point, computation, is not the name of a physical process, in a way that, for example, digestion, or vision, human vision is the name of a physical process. Computation names a formal, or syntactical process, that can be realized in an indefinite range, of different computer hardwares, but that multiple realizability, is a clue that we're not talking about something in physics. Computation does not name something is physics, or chemistry, and that's the failure of an analogy between computation which is not a physical process, and phenomena like electromagnetic radiation which actually are.

Well I'll mention one more of these numerous criticisms; this one is due to Jerry Fodor. Fodor said look, the whole point about computers is you gotta have a mechanical implementation of the program, but in the Chinese Room you are cheating, cause the guy carrying out the steps in the program, is an actual guy, an actual human being, an actual Searle, going though the steps in the program, and that's cheating. Cause the brain, when it is acting computationally, it does it without the benefit of anybody in there carrying out the program consciously. So if you have a conscious implementation of the program, that's not an implementation of the program. Now when I first read that, I really didn't know whether to laugh or cry for Alan Turing. Because remember, this is Turing's definition that we're using here, and what Fodor, in effect, is saying is, that Turing really didn't know what a Turing machine was, and that's a very nervy thing to say. You see, here's the irony of this, the word computer has changed its meaning. When Turing wrote his original work, computer meant, person who computes. The way runner means someone who runs, and when Turing talked about a computer, he meant a person who computes. And that's why his famous article is not called, Computers and Intelligence, because that would be about people, it's called, Computing Machinery and Intelligence, and computing machinery was opposed to computing. Now in the way in which languages have changed, we now call computing machinery, we call them computers. But in the old days the computers were like runners, or writers, ah-it didn't mean machine that wrote or write, it meant a person who ran, or write, and what Fodor has done is in effect, try to take our changed definition of computer, ah so computer now means machine, to say a person computing isn't really a computer. But it's a totally arbitrary move, the definition of computation, which we got from Alan Turing, says; any appropriate-syntactical-process, is a case for computation, because that's how computation is defined.

I mention that example cause I think it is a symptom of the desperation, of a lot of the people involved in this field, and as I said earlier, I don't fully understand, the passion, that this issue arouses. I am amazed both in my own case, and in the case of other people who have attacked the computational model of the mind, because I think that it’s scientifically, and philosophically, inadequate, but the extremity of the attacks, and the passion that this arouses, suggests the conviction that people have, that we’re all computers, is a bit like a religious conviction. It’s the feeling like there’s something, very important, maybe even more important than our research grants, is gonna be lost, if it turns out that we’re not computers.

Anyway, I could continue this list of criticisms of the Chinese room, but I wanna open the discussion now further. I mean we’ve already hinted, at some of these larger issues and let me know shift to what I think is the issue that has been underlying much of this debate. Remember Descartes,