PH1102E/GEK1067

Week 11

Sims

I. Sims vs. Naturals

Suppose you have good reasons to believe that spacetime contains an astronomical number of **computer-generated mental lives** that are indistinguishable from ordinary human mental lives. They are indistinguishable from ordinary human mental lives, in the sense that they consist of the same kinds of thoughts, feelings, moods, emotions, etc. as ordinary human mental lives like yours and mine. So it's impossible to tell by introspection ("from the inside") whether your mental life is, or is not, computer-generated.

Let's define a "humanlike mental life" as a mental life (computer-generated or not) that is, in the sense just explained, indistinguishable from an ordinary human mental life. (So, your own mental life is a humanlike mental life.) And let's define a "Sim" as someone with a computer-generated (CG) humanlike mental life:

A **Sim** is someone with a CG mental life that is subjectively indistinguishable from an ordinary human mental life.

Finally, let's define a "Natural" as someone with a non-computer-generated humanlike mental life:

A **Natural** is someone with a humanlike mental life that is not computer-generated.

Subjectively, the life of a Sim is indistinguishable from the life of a Natural. If you were a Sim, nothing about the quality of your conscious thoughts or experiences would tip you off to that fact. From the inside, the life of a Sim is no different from yours or mine.

Suppose that you have good reasons to think that there are so many Sims in the universe that the ratio of non-CG lives to CG lives is extremely small:

$$\frac{\text{number of Naturals}}{\text{number of Sims}} = \text{almost zero}$$

Now, even though you have (we are supposing) good reasons to believe that almost all humanlike sentiences that exist in spacetime are Sims, let's suppose that you have no idea which region of spacetime contains the Sims. For all you know, the Sims might all exist in the past, or all in the present, or all in the future; for all you know, some might exist in the past and some in the present and some in the future; etc. All that you can be sure about (we're supposing) is that if you add up all the Sims that ever exist---past, present, or future---the total number of Sims vastly exceeds the total number of Naturals.

If all this is true, how confident should you be that your own mental life is not computer-generated? That is, how confident should you be that you are a Natural, rather than a Sim?

The answer, apparently, is: not very confident at all.

Imagine a vast urn filled with billions of marbles. Almost all of the marbles are black; the rest are white. Your recently deceased great-aunt has bequeathed one of the marbles to you in her will, but you don't know which one. All you know is that one of the marbles in the urn now belongs to you. How confident should you be that your marble is a white marble? Answer: not very.

Think of the urn as representing spacetime, the black marbles as representing Sims, and the white marbles as representing Naturals. If spacetime contains astronomically more Sims than Naturals, then you should have very little confidence that your own mental life is a Natural mental life. In other words: you should be almost certain that you are a Sim.[1]

So: if we have good reasons to believe that spacetime contains vastly more Sims than Naturals, then we should be almost certain that we are Sims.

But: do we have good reasons to believe that spacetime contains vastly more Sims than Naturals? Nick Bostrom suggests that we may!

II. The Simulation Argument

According to Bostrom, we good reasons to believe that Sims vastly outnumber Naturals, provided that we have good reasons to believe: (1) that human civilization at some point in time acquires the technological capability to run "ancestor simulations," and, (2) that any civilization that acquired the capability to run ancestor simulations would run a lot of ancestor simulations.

Now, Bostrom himself doesn't assert either (1) or (2). His position is that there's at least a good chance that (1) is true, and at least a good chance that (2) is true. But let's see what happens when we use (1) and (2) as the premises of an argument, as follows:

(1) At some point in time, human civilization acquires the technological capability to run ancestor simulations.

(2) Any civilization that acquired the capability to run ancestor simulations would run a lot of ancestor simulations.

(3) So, at some point in time, human civilization runs a lot of ancestor simulations. (follows from 1 and 2)

[1] You can also think of it this way. Suppose that God has the Universe before Him in its spatial and temporal entirety, and that He sees that spacetime contains far more Sims than Naturals. Now God announces that he is going to give an eternal reward to all the Naturals. How excited should you be? Well, if you know that spacetime contains far more Sims than Naturals, you shouldn't be very excited at all.

(4) But if at some point in time human civilization runs a lot of ancestor simulations, then spacetime contains a vast number of Sims---vastly more than the number of Naturals that it contains.

(5) So, spacetime contains vastly more Sims than Naturals. (follows from 4 and 5)

This argument contains three underived premises: (1), (2), and (4). If you accept these premises, you must accept the conclusion of the argument, (5), and therefore accept that in all likelihood you are a Sim.

Premise (1)

This premise states that there is some time at which humanity acquires the ability to run "ancestor simulations." What is an ancestor simulation?

An ancestor simulation is a detailed computer model of human history, or some part of human history, or some variant on human history. It's like a computer model of the Solar System, only vastly more complex.

Imagine a cubical region of space containing the Earth, including the Earth's atmosphere and all the Earth's inhabitants (plants and animals, including human beings). Now imagine that we divide the cube into a fine 3D grid. Each cubical cell or "block" of the grid is microscopically small---say, a cubic nanometer (that's around one one-thousandth the size of a brain cell).

To give a detailed description of human history over the past million years, it would be enough to state, for each block in the 3D grid, and for each moment of time over the past million years, what that block contained at that moment. (By a "moment," I mean some very brief period of time---say, a nanosecond.) That is to say: you could describe the past million years of human history by specifying the contents of each block in the 3D grid at each moment over the past million years.

Since the blocks are so small, there's only so much that can fit inside one of them at any given time, and only so many properties that the contents of each block can exhibit: simple physical or chemical properties (of mass, charge, spatial orientation, etc). So, if we wanted to give a detailed description of the past million years of human history, it would be enough to make a big chart, with one row for each 3D block, and one column for each moment of time (over the past million years), and fill each cell of the chart with a list of the properties possessed by the stuff occupying the corresponding block at the corresponding moment (if the block was empty at a given moment, we leave the corresponding cell of the chart empty). Letting "t_1," "t_2," "t_3," etc. stand for successive moments of time, and "P_1," "P_2," "P_3," etc. stand for various simple physical properties, the chart might look something like this (of course, this would only be a small part of the chart; the whole thing would contain a huge number of rows and columns):

	t_1	**t_2**	**t_3**	**t_4**	**t_5**	**t_6**	**t_7**	**t_8**	**t_9**	· · ·
Block 1		P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}		P_7 P_{26} P_{781}	P_7P_{26} P_{781}					
Block 2		P_{271} P_{655} P_{879}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}			P_{10} P_{130} P_{572}	P_{287} P_{383} P_{709}			
Block 3				P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}	P_1 P_3					
Block 4	P_2 P_{863} P_{947}	P_{16} P_{106} P_{303} P_{342} P_{789} P_{853}	P_2 P_{863} P_{947}	P_{16} P_{106} P_{303} P_{342} P_{789} P_{853}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}	P_{16} P_{106} P_{303} P_{342} P_{789} P_{853}	P_2 P_{863} P_{947}	P_{16} P_{106} P_{303} P_{342} P_{789} P_{853}	P_2 P_{863} P_{947}	
Block 5	P_{123} P_{139}	P_{123} P_{139}	P_{420} P_{932}	P_{420} P_{932}	P_{701} P_{745}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}			P_{42} P_{101}	
Block 6		P_{469} P_{506}		P_{12} P_{380} P_{717}		P_{15} P_{119} P_{367} P_{749}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}	P_1 P_{213} P_{569} P_{811}		
Block 7							P_3 P_7 P_{12}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}	P_1 P_{213} P_{569} P_{811}	
Block 8	P_{55} P_{102} P_{277} P_{279} P_{500} P_{796} P_{889} P_{973}			P_2 P_{27} P_{81} P_{165} P_{702}			P_1	P_{33}	P_1 P_{27} P_{81} P_{165} P_{402} P_{411} P_{627} P_{801} P_{805}	
Block 9	P_2 P_8	P_2 P_8	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	P_4 P_7 P_{29} P_{170} P_{823}	
⋮										

Now suppose that we want to program a computer, based on this Big Chart. Recall that a computer is essentially just a big array of microscopic on-off switches. Our goal is to program the computer so that changes in the positions of its microswitches over time mirror the changes that took place over the past million years within the big cube of space containing the Earth.

To make things more concrete, let's suppose that there are exactly 1000 simple physical nano-properties (P_1 through P_{1000}).[2] Then we're going to need a computer with 1000 times as many microswitches in it as there are cells in the Big Chart. Each set of 1000 switches corresponds to a cell in the Big Chart. For example, there's a set of 1000 switches corresponding to the highlighted cell (Block 6, t_4); call this Set 6. This cell indicates that Block 6 of the 3D array contained, at time t_3, something having properties P_{12}, P_{380}, and P_{717}. (Maybe at t_3 Block 6 contained a particle of matter that was made of gold (P_{12}), had a mass of twelve nanograms (P_{380}), and a temperature of sixteen degrees Celcius (P_{717}). So we program our computer in such a way that at the third "tick" of the computer clock, all of the microswitches in Set 6 are turned off except for switches #12, #380, and #717.

There are many different programs we can run on the computer. Each program tells the computer to set certain of its switches On and the rest Off with each successive tick of the computer's internal clock. Each successive internal state of the computer---each totality of On/Off settings of the switches---corresponds to a column of the Big Chart. And each column of the Big Chart corresponds to the nanoscale state of the Earth (and everything on it) at a certain moment of time. So, each successive internal state of the computer corresponds to the nanoscale state of the Earth (and everything on it) at a certain moment.

A program that instructs the computer to run through a certain sequence of internal states corresponds to a certain version of the Big Chart---a certain filling-in of the Chart's cells with P_n values. And any given version of the Big Chart corresponds to a certain version of the history of the Earth over the past million years. It is in this sense that the computer **simulates** or **models** a million-year period of Earth history.

What makes our computer an ancestor simulator is that the million-year period of Earth-history it simulates includes a large chunk of the history of the human race. By running a detailed simulation of the past million years of Earth-history, we automatically run a detailed simulation of the past million years of human-body history, including the past million years of human-brain history.

So much for what an ancestor simulator is; why should we believe (as Premise (1) states) that at some point in time human civilization acquires the technological capability to run ancestor simulations?

Well, for one thing, we already have the technical know-how to build computers powerful enough to run ancestor simulations. In fact, we already know how to build a computer powerful enough to simulate the entire history of the human race in a millionth of a second. (That is, we know how to build a computer that can change the On/Off configuration of its microswitches so rapidly that it can run through all the configurations represented by successive columns of the Big Chart in a millionth of a second.) It's just that the cost of

[2] The number of basic physical properties could actually be much larger than this, without affecting the argument.

building such a computer with presently available technology is prohibitive. But there is a long-standing historical trend (sometimes referred to as “Moore’s Law”) by which the cost of building a computer with a given amount of computational power (i.e., capable of a given number of “flops” a.k.a. changes of On/Off-switch-configurations per second) declines by about half every two years or so. If this trend continues, it will be affordable to build computers powerful enough to run ancestor simulations within the next two or three hundred years.

Of course, you need more than just a powerful computer to run an ancestor simulation. You also need to program the computer in the right way. By far the most challenging aspect of this programming task is programming a computer to simulate the nano-scale activity that takes place in human brains, human brains being the most complex systems (by far) on Earth. But with advances in neuroscience, and improvements in computer-assisted programming, it does not seem far-fetched to suppose that within the next couple hundred years, we’ll have the knowledge required to write an ancestor-simulation program. (Historically, it is hardware-development that has limited software-development---scientists always have a use for more powerful hardware.)

So <u>if</u> the human race can avoid extinction or widespread social catastrophe for the next thousand years, it seems quite likely that humanity will develop ancestor simulation technology. So we should accept Premise (1) if we are reasonably optimistic about the survival prospects for human civilization. Well, for the sake of argument, let’s be optimists, and move on to Premise (2).

Premise (2)

Premise (2) states that any civilization that acquired the capability to run ancestor simulations would run a lot of ancestor simulations. This seems pretty uncontroversial. After all, if <u>we</u> had the capability to run ancestor simulations, wouldn’t we run a lot of them? Think of all the potential applications of this technology: virtual history tours, highly realistic history-based virtual-reality gaming, testing historical hypotheticals (“what if the Nazis had developed the atom bomb before the Allies?”), etc.

Premise (4)

Premise (4) states that if at some point in time human civilization runs a lot of ancestor simulations, then spacetime contains a vast number of Sims---vastly more than the number of Naturals that it contains.

This premise is based on the assumption that **<u>simulated</u> brain activity generates <u>real</u> thoughts and experiences (sensations, feelings, etc.).** Why should we accept this?

According to some philosophers of mind---the dualists---there is more to having conscious thoughts and experiences than merely possessing a brain (or other physical system) that operates as a complex network of interdependent elements. Dualists maintain that there is no outright contradiction in the idea that a system isomorphic to a normally functioning human brain---such as a computer system that simulated brain activity at a detailed (e.g., nanoscale) level---might fail to produce any conscious thoughts or experiences.[3]

However, even most dualists think that we have good reasons to believe that *as a matter of fact*, if you were to create a detailed simulation of the activity of a normal human brain, you would get genuine thought and experience as a result.

The dualist David Chalmers argues for this with something he calls the "fading qualia" argument. Suppose that as you're watching TV, some doctors are somehow secretly replacing the neurons in your visual cortex (the part of your brain responsible for visual imagery) with computer-chips that perform the same function as the neurons they replace. They "perform the same function," in the sense that they have the same effects on, and are affected in the same ways by, the neurons to which they are connected. By the end of the procedure, the doctors have replaced all of your natural neurons with silicon prostheses.

Question: when the procedure is over, do you have visual experiences? Or are you blind and incapable incapable of having any sort of visual imagery? But remember that in terms of *function*---input and output---nothing has changed in your brain. Your "wiring" remains the same, just using different kinds of wires and circuits (synthetic instead of organic). So, when we *ask* you at the end of the procedure whether you can see, you'll say "Yes" with perfect sincerity. And when you get up off the sofa to get a drink from the refrigerator, you'll navigate your way as smoothly as ever. In a word: all of your *behavior*, including your verbal and cognitive behavior, will remain the same as it was before the surgery. But how strange all this would be, if you were now completely without conscious visual experience!

There are also more general, methodological reasons to suppose that any process isomorphic to a given conscious process is itself conscious. If something more than isomorphism with a conscious process is required to generate conscious thought and experience, what more is this? Unless we are prepared to attribute occult powers to flesh---powers irreproducible in any other medium---it is hard to see how we can fail to take isomorphism with our own neural processes as decisive evidence of conscious thought and experience like ours.

The upshot of which is that conscious mental lives just like our own could, in theory, be conjured up in an immensely powerful computer that simulated the kind of nanoscale

[3] Two processes are "isomorphic" if they have the same number of parts (parts being counted separately accordingly as they play separate roles in the process), and if the relationships among these parts in the one process mirror the relationships among the parts of the other process. A "part" of the process can be an event that occurs during some stage of the process, or an object that is involved in the process, or a state of such an object at some point in the process. The processes that take place in one of Bostrom's ancestor simulations are isomorphic to the nano-processes that occur in human bodies and their environment.

physical activity that takes place in a human brain. Such a computer simulation would, if Chalmers is right, generate conscious mental lives *just like ours*. And if that is correct, we must accept Premise (4).

III. Consequences of the Simulation Argument

If the Simulation Argument is sound, it is very likely that we are Sims. Does this mean that it is also very likely that the world we take ourselves to inhabit is an illusion?

We take up this question in our next lecture. For now, I simply point out one definite implication of the idea that we are Sims, which is that we have Simulators. If, as the Simulation Argument purports to prove, we are living in a computer simulation, then there are beings, or a Being, who has set the parameters of the simulation, and who literally controls our lives and our world.

In this way, the Simulation Argument, like the Cosmological Argument, purports to establish the existence of a kind of transcendent force or power as the source of the world as we know it. It would also be interesting to consider the Argument from Evil in light of the Simulation Argument, but that is a conversation for another day.