1. Are you living in a computer simulation?

Are you living in a computer simulation?

This question is implicit in films like *The Matrix* or *The Thirteenth Floor*, both of which involve characters who discover that the world they inhabit is a simulation. The question, "Are you lving in a simulation?", is analogous to Fermi's famous question about "Where are the aliens?"

Fermi's question motivated the famous 1960 Drake Equation, which recast the question in probabilistic terms. That reframed the debate and has been the basis for discussion ever since.

In 2003, Nick Bostrom did the same thing with respect to the question, "Are you living in a simulation?" His approach raised fundamental philosophical and metaphysical issues that challenged long-standing ideas. It also provided a novel, probabilsitic way of framing the question.

To frame the question, he starts with two explicit assumptions.

- **(1)** We will (eventually) be able to produce *simulated people*, i.e, artificial intelligences with the ability to process information in a way thats indistinguishable from human intelligences.

- **(2)** we will be able to produce a *simulated reality* of sensory inputs for these simulated people, and do so in such a way that the simulated people would be unable to determine that they were in a simulation.

There is a third implicit assumption

- **(3)** if intelligent beings *could* run such a simulation, they *would*. In fact, the utility of such a simulation would be so great, that it's inevitable there would be many such simulations.

In support of the third assumption, it's worth noting that humans have been consciously "running simulations" at least since Plato first wrote about forms. Fictional simulations–think Homer–are even older. Today we use sophisticated simulations to help understand everything from the Big Bang, to the stock market, to the human genome. Fictional simulations like two mentioned above abound. We're nowhere near the abilities posited in Bostrom's two assumptions, but we'e at the point where they are are at least plausible.

The basic idea of Bostrom's analysis is implicit in assumption (1): that there are two kinds of people. There are "real people" and "simulated people." If that's the case, then each group has a population size.

$$\boldsymbol{N_{Sim}} = \text{total number of simulated people}$$

$$\boldsymbol{N_{Re}} = \text{total number of real people}$$

From this, it follows that the total number of "people" is

$$\boldsymbol{N_{Tot}} = \text{total number of people, simulated and real, i.e.}$$

$$\boldsymbol{N_{Tot}} = \boldsymbol{N_{Re}} + \boldsymbol{N_{Sim}}$$

Now select a person at random. Based on the two assumptions, we don't know if that person is real or simulated. What are the chances the person we select is real? Well, that's easy. It's

$$\frac{N_{Re}}{N_{Tot}}$$

or

$$\frac{N_{Re}}{N_{Re} + N_{Sim}}$$

Now, it's the real people who are–at least initially–creating the simulated people, so there's a relatsionship between N_{Re} and N_{Sim}. At the simplest level, the relationship might be linear

$$N_{Sim} = k \times N_{Re}$$

where k represents the number of simulations per real person. However, it gets more complicated, since there might be simulated people in simu-

lated reality running their own simulations–this is exactly the situation in *The Thirteenth Floor.*

Related issues have to do with how much computing power (measured, say, in operations per second) it takes to run a simulation. That was ultimately the heart of Bostrom's analysis, since it turns out the number of possible simulations–and hence the number of simulated people–is huge. In fact, assuming we reach the point where we've achieved assumptions (1) and (2), N_{Sim} is a **colossolly** huge number compared with any estimate of N_{Re}.

Thus the, chances of a randomly selected person being real, given by the fraction

$$\frac{\boldsymbol{N_{Re}}}{\boldsymbol{N_{Re}} + \boldsymbol{N_{Sim}}}$$

is essentially zero since the denominator is a huge number.

Consequently, if we–or any collection of intelligent beings–ever reach the point where we've achieved (1) and (2), the chances a person selected at random is real is essentially zero. Conversely, then, the chances that *you are simulated person* is virtually certain.

In fact, if we are living in a simulation, it's one that was created by real people who *have* achieved (1) and (2).

It looks like this doesn't provide an answer after all, but it does, at least in a way. If any collection of intelligent beings ever achieves (1) and (2), then the analysis shows we are almost certainly living a simulation created by these beings. The alternative is that *no collection of intelligent beings ever achieves (1) and (2).*

Since we've agreed that (1) and (2) are at least plausible outcomes, that leaves us with one of two binary choices:

- **(A)** Either intelligent beings always self-destruct before achieving (1) and (2) (which implies we can't be living in simulation since they can't exist); or

- **(B)** Intelligent beings *can* achieve (1) and (2), which implies we ARE living in a simulation.

As with Drake's equation, Bostrom's involves parameters that are unknown and hard to estimate. Small changes in the assumptions can lead to massive changes in things like the estimates for $\boldsymbol{N_{Sim}}$ which, in turn, are critical to the two main conclusions (A) and (B) above. These conclusions can and have been debated. But the point here is that Bostrom reframed the argument in a novel way that has influenced debate ever since.

N. Bostrom, Are you living in a computer simulation?, Philosophical Quarterly 57(211): 243-255 (2003)