Class Five: The Simulation Argument

Philosophy and Science Fiction - Ryan Simonelli

October 12, 2022

1 The Simulation Agument, Weak and Strong

- **Bostrom's Disjunction:** The conclusion of Bostrom's argument (what I'll call the "Weak Simulation Argument") is that one of the following three propositions is true:
  1. The human species is very likely to become extinct before reaching a "posthuman" stage.
  2. Any posthuman civilization is extremely unlikely to run a significant number of simulations of its evolutionary history (or variations thereof)
  3. We are almost certainly living in a computer simulation.
- **A Bit of Logic:** From the proposition "p or q or r" and the proposition "not-p and not-q," the proposition that r follows.
- **The Strong Simulation Argument (as I'll call it):**
  - Bostrom's Disjunction: (1) or (2) or (3)
  - not-(1) and not-(2)
  - So, (3)
- **Clarifying the Positions at Play:**
  - What Bostrom calls "The Simulation Argument" is what I'm calling the *Weak* Simulation Argument. People other than Bostrom have often used the term "The Simulation Argument" to speak of the *Strong* Simulation Argument.
  - The "Simulation Hypothesis" is the conclusion of the Strong Simulation Argument, the claim that we're (almost certainly) living in a simulation.
  - Bostrom himself doesn't endorse the Strong Simulation Argument, only the Weak One. He thinks propositions (1), (2), and (3) are all equally plausible. Many people who are persuaded by Bostrom's considerations, however, are drawn to the Strong Simulation Argument.

2 Some Assumptions

- **Substrate-Independence/Multiple Realizability of Consciousness:** A common idea in philosophy of mind: mindedness, and more specifically consciousness, can arise out of processes constituted by a variety of different physical substances.
  - **Intuiton:** You could replace my braincells one by one with silicon chips that functions just as a brain cells do. At the end of the process, I'd have an entirely silicon brain, but, plausibly, I'd still be conscious.
  - **Consequence:** If this is so, then the processes of a simulated brain, implemented in a powerful computer, could result in conscious experiences.
- **Realistic Simulations of the Universe Are Technologically Possible:**

- **Full Brain Simulations Are Not Out of Reach:** The number of computations per second carried out on a theoretically possible computer of a feasible size exceeds the number of computations per second carried out by the human brain by several orders of magnitude, making it possible for billions of brains to be completely simulated.
- **Shortcuts Will Be Necessary:** Clearly, we could not simulate the entirety of a universe as complicated as this one within this very universe. But we plausibly could simulate a subjectively indistinguishable universe by only fully rendering and fixing things as necessary upon observation or interaction.

3 Bostrom's Official Argument

- **Deductive vs. Probabilistic Arguments:** Most of the arguments that we've considered so far are *deductive*, in that, if you lay them out formally, the conclusion follows from the premises as a matter of *deductive logic*. Bostrom's argument instead uses *probability theory* to show that a certain relationship obtains between three propositions.
- **Some Notation:** Let us write
  - f_p for the fraction of human-level technological civilizations that survive to reach a posthuman stage.
    - So, for instance, if 50% of civilizations at our level of technological development survive to a posthuman stage so as to have massive computational power at their disposal, then $f_p = .5$
  - $\bar{N}$ for the average number of ancestor-simulations run by a posthuman civilization.
    - Note that this factors in the civilizations that don't run any simulations.
  - $\bar{H}$ for the average number of individuals that have lived in a civilization before it reaches a post-human stage.
  - f_{sim} for the fraction of individuals with human-like experiences that live in simulations.
    - So, for instance, if $f_{sim} = .99$, then 99% of individuals with experiences that are like ours live in simulations.
- **Bostrom's Basic Equation:**

$$f_{sim} = \frac{f_p \bar{N} \bar{H}}{(f_p \bar{N} \bar{H}) + \bar{H}}$$

  To unpack this a bit, the top part of the fraction equates to the average number (for a given civilization) of individuals with human-like experiences that live in simulations, and the botom part of the fraction equates to the average number (for a given civilization) of individuals with human-like experiences that either live in simulations or not.
- **An Example:** Suppose, for instance, we have the following facts:
  - 50% of human-level civilizations of survive to a posthuman stage
  - A posthuman civilization runs, on average, 1000 simulations
  - The average the average number of individuals who live in a civilization before it reaches a posthuman stage is 200 billion (for reference, currently, there is an estimated 117 billion people who've ever lived).

  Then, for an average civilization, the number of individuals with human-like experiences who will ever have lived is .5 x 1000 x 200 billion, which gives us (if I did my math right) 100 trillion. The fraction of people who are simulated, then, would be 100 trillion over 100 trillion, 200 billion, or about 0.998.
- **Tweaking the Equation:** Using a bit of high school math, we can factor out the $\bar{H}$ and simplify as follows:

$$f_{sim} = \frac{f_p \bar{N} \bar{H}}{(f_p \bar{N} \bar{H}) + \bar{H}} = \frac{\bar{H}(f_p \bar{N})}{\bar{H}((f_p \bar{N}) + 1)} = \frac{f_p \bar{N}}{(f_p \bar{N}) + 1}$$

- **Substituting:** Writing f_I for the fraction of posthuman civilizations that are interested in running (and so do run) ancestor-simulations and $\bar{N}_I$ for the average number of simulations run by such a civilization, we have that $\bar{N} = f_I\bar{N}_I$. So, substituting for $\bar{N}$, we get the following equation:

$$\frac{f_p f_I \bar{N}_I}{(f_p\ f_I\bar{N}_I) + 1}$$

- **The Official Disjunction:** Given that $\bar{N}_I$ is very large (plausibly, at least in the thousands), it follows that at least one of the following propositions is true:
  1. $f_p \approx 0$
  2. $f_I \approx 0$
  3. $f_{sim} \approx 1$
- **The Bland Indifference Principle:** If (3) is true, given that we have no reason to suppose our experience is distinctive of a non-simulated being rather than a simulated one, our credence that we are a simulated being should be close to 1. That is, we should be almost certain that we are a simulated being, living in a computer simulation.

4 Considering the Plausibility of the Three Disjuncts

- **First Disjunct:** The fraction of human-level civilizations reaching a post-human stage is near zero.
  - Quite pessimistic if we accept it!
  - In certain respects, technological advancement seems to make us less likely to go totally extinct.
    - It won't be that long until we're a multi-planetary civilization, so a catastrophe on one planet wouldn't necessarily spell the end of our civilization.
    - NASA just intentionally crashed a ship into an asteroid to deflect it, and this sort of technology could be used to avoid natural disasters.
  - In other respects, technological advancement seems to make us more likely to go totally extinct.
    - Many people think that global warming, caused by humans, is a serious existential threat.
    - Nuclear war seems like a real possibility these days, and this could
  - But note, even if we think that about half the civilizations like ours will completely destroy themselves (and even that seems like a somewhat pessimistic estimate), that's still too far high of a percentage for the truth of (1). For (1) to be true it must be the case that *almost all* such civilizations destroy themselves. That seems somewhat implausible to me.
  - **The Fermi Paradox, A Related Issue:** It seems that, if posthuman societies *do* exist, then such civilizations will have intergalactic travel. If that's the case, however, we should have seen them by now!
    - One conclusion to draw from the fact that we *don't* see them could be that they all destroy themselves before they get to that level.
    - It could also be, however, that they simply don't interfere with us, just as we're not supposed to interfere with the wildlife in nature preserves.
- **The Second Disjunct:** The fraction of posthuman civilizations who opt to run ancestor simulations is near zero.
  - If we consider ourselves as an example, this seems implausible.

        - Consider video games. We are continually trying to make better and better simulations, with better AI for entertainment purposes. If we could make a hyper-realistic version of the *Sims*, would we not do it?
        - Beyond purposes of entertainment, it seems like there'd be great scientific value in running such simulations, to see how various features of ourselves and our world could have evolved and emerged.
    - Perhaps there's ethical reasons why a posthuman civilization would not run such simulation.
        - But even if we suppose that posthuman civilizations will be universally ethical, it's not obvious that there's anything unethical about running such a simulation insofar as one does so ethically (unlike the way I play Sim City).
        - Wouldn't bringing billions of meaningful conscious lives into existence be, on the whole, a good thing?
    - **An Interesting Argument:** Bostrom considers the possibility of simulations being created *inside* simulations. This, however, poses serious problems for the base level simulators—a planet-sized computer trying to simulate a planet-sized computer running a simulation would presumably make the computer crash. Bostrom proposes, in that case, that "we should expect our simulation to be terminated when we are about to become posthuman," (253). Perhaps then, posthuman civilizations, acknowledging this possibility, don't run simulations out of fear that they themselves are simulations that might crash the program if they do. (There's an odd sort of self-defeating character to this argument, if you think about it for a bit, since its based on the assumption that it's likely we're in a simulation, but it undercuts the claim that it's likely we're in a simulation.)
- **The Third Disjunct:** The fraction of beings with human-like experiences who live in simulations is near one (and so we are almost certainly living in a simulation).
    - Most philosophers would regard this possibility with horror. According to both Chalmers and Bostrom, however, it's really not *that* big of a deal.
        - It's not a radical skeptical hypothesis, as most philosophers have supposed. Even if we're simulated beings, most of our beliefs are true—they're just about virtual objects rather than non-virtual ones.
        - It doesn't make life meaningless, as Nozick supposes. One's still doing things and being a specific sort of person, if one's a simulated beings—it's just that one is doing things in a virtual world and one is a virtually embodied person.
    - Still, it's a *pretty dang big deal*.
        - Both Chalmers and Bostrom compare the simulation hypothesis to a *theological* hypothesis, albeit one that's compatible with naturalism. At least if one is an atheist, the idea that there are god-like entities—the creators and maintainers of the simulation—who are "omnipotent," in that they can intervene in our world and do whatever they want, and "omniscient" in that they can monitor everything that happens is startling to say the least.
        - If the simulation hypothesis is true, then we have reason to be concerned about the possibility of the simulators turning off the simulation, or about something happening in their universe (over which we have no control) that destroys the computer on which the simulation is running. In general, it seems that our existence is fragile in a way that is utterly out of our control, and that's a rather scary prospect.