A PATCH FOR THE SIMULATION ARGUMENT

Nick Bostrom
Future of Humanity Institute
Faculty of Philosophy & Oxford Martin School
University of Oxford

Marcin Kulczycki
Institute of Mathematics
Faculty of Mathematics and Computer Science
Jagiellonian University

[Published in: *Analysis*, Vol. 71, No. 1 (2011): 54-61]

www.simulation-argument.com

Abstract

This article reports on a newly discovered bug in the original simulation argument. Two different ways of patching the argument are proposed, each of which preserves the original conclusion.

The bug

An earlier paper by one of us (N.B.) argues that, having accepted some plausible assumptions, one must conclude that at least one of three propositions is true:

(1) The human species is very likely to go extinct before reaching a posthuman stage[i]
(2) The fraction of posthuman civilizations that are interested in running a significant number of ancestor simulations is extremely small.[ii]
(3) We are almost certainly living in a computer simulation.[iii]

This paper has generated several commentaries from the philosophical and scientific community and has drawn considerable interest from the wider public.[iv]

What has so far passed unnoticed is a mathematical *non sequitur* in the original paper. At the heart of the argument is a formula for calculating f_{sim}, the fraction of all observers in the universe with human-type experiences that are living in computer simulations:

$$f_{sim} = \frac{pNH}{pNH + H}$$

Here p is the fraction of all human-level technological civilizations that manage to reach a posthuman stage, N is the average number of times a posthuman civilization runs a simulation of its entire ancestral history, and H is the average number of individuals that have lived in a civilization before it reached a posthuman stage.[v]

In order to see the problem with this formula, imagine a universe in which only two civilizations developed, out of which the first consisted of $3X$ beings and ended without reaching a posthuman stage, while the second reached a posthuman stage after X beings had lived in it, at which point it ran N simulations of its ancestral history. The above formula reports that $f_{sim} = N/(N+2)$ while in truth the fraction is $N/(N+4)$.

By choosing different numbers, more extreme differences can be obtained. Consider the following model: There is one civilization in which $99(99N-1)X$ people lived and which never reached a posthuman stage. In addition, there are 99 civilizations that reached a posthuman stage after X people lived in each of them. Assume that each of those 99 civilizations run N simulations of its entire ancestral history. Simple calculations then show that:

(A) The fraction of human-level civilizations that reached a posthuman stage was 99% .
(B) The fraction of posthuman civilizations that decided to run ancestor simulations was 100% .
(C) A full 99% of all persons lived non-simulated lives.

This result would seem to suggest the possibility of the three propositions in the central tripartite disjunction of the simulation argument all being simultaneously false, thus undermining the argument's conclusion.

The vulnerability

The basic problem can be simply stated. Let us say that a civilization starts out unable to create ancestor simulations (call this the "pre-posthuman" phase) and possibly later becomes able to create such simulations (in a "posthuman" phase). Now, if those civilizations that eventually reach a posthuman phase have *unusually brief pre-posthuman phases* compared to other civilizations, then—since the ancestor simulations only cover the pre-posthuman phase—it could happen that most pre-posthuman observers live outside simulations even if most pre-posthuman civilizations eventually become posthuman, and even if each posthuman civilization runs several ancestor simulations. This is the underlying vulnerability that can lead to violations of the tripartite disjunction.

We will now present two alternative ways of patching the simulation argument to remove this vulnerability. The two patches are independent of one another and individually sufficient.

The first patch

The first way to patch the argument starts by noting the empirical claim, argued for in the original paper, that a posthuman civilization would have the capability to run *an astronomical number* of ancestor simulations, even using only a tiny fraction of its computational resources for that purpose. Given this, we need only introduce a very weak assumption to the effect that the typical duration (or more precisely, the typical cumulative population) of the pre-posthuman phase *does not differ by an astronomically large factor* between civilizations that never run a significant number of ancestor simulations and those that eventually do. For example, in an appendix we show how by assuming that the difference is no greater than a factor of one million we can derive the key tripartite disjunction. (If the empirical estimates in the original paper are in even the right ballpark, this assumption could be weakened by many additional orders of magnitude.)

To appreciate the empirical plausibility of this added assumption, consider that if, for instance, civilizations that run a large number of simulations rarely had much fewer than 100 billion people living in their pre-posthuman phases (which is about the number of human beings that have already lived on Earth today at a time when we have not yet reached a posthuman phase) then civilizations that never ran a significant number of ancestor simulations would each need to have an average cumulative population of over 100 million billion pre-posthumans in order for the assumption to fail. Even if the world population reaches, and remains at, 20 billion, this would allow for five million pre-posthuman generations—extended over some 100 million years. One would think that 100 million years is ample time for a species like Homo sapiens to either go extinct or develop posthuman levels of technology. (And again, we could increase this bound by *many* orders of magnitude if we weaken the requisite empirical assumption as much as possible.)

The second patch

The second way to patch the argument is by taking into account information about our own place in history. We may be uncertain about whether the world we experience is simulated or not; and conditional on it being simulated, we may be uncertain about how many simulations have been run before ours: yet we still know something about our position *within* our world. For example, we know that in our history:

- The human species evolved some hundred thousand years ago.
- Some 100 billion people have been born thus far.
- An industrial revolution took place a couple of hundred years ago.
- The first 1 MHz processor was created just under forty years ago.

We can bring such knowledge to bear when we assess the probability that we are in a simulation. We do this by asking where most observers *with our kinds of*

experiences live, conditional on (1) and (2) being false. Given the background assumptions stated in the original paper, there would, if (1) and (2) are false, be many more simulated histories than non-simulated histories. Now, even if it were the case that each of the non-simulated histories contained far more people than each of the simulated histories (perhaps because the pre-posthuman phase lasted far longer for those civilizations that never produced simulations), this would not need preclude it being true that most people with our kinds of experience exist in simulations.

What this patch needs in order to work is that we have some empirical indexical evidence E such that it is plausible to assume that it satisfies the following conditions. (Here, E is some centered proposition. An *E-observer* is an observer about whom E is true.)

(i) In a substantial fraction of those pre-posthuman histories that end up running (significant numbers of) ancestor simulations, there is some E-observer.

(ii) Let $H_s(E)$ be the average number of E-observers among those pre-posthuman histories that contain some E-observer and that end up running (significant numbers of) ancestor simulations. Let $H_n(E)$ be the average number of E-observers among those pre-posthuman histories that contain some E-observer and that do not end up running (significant numbers of) ancestor simulations. It is *not* the case that $H_n(E)$ is vastly greater than $H_s(E)$.

(iii) There is no defeater, i.e. we have no *other* information that enables us to tell that we are not in a simulation. (A defeater could be some more specific centered proposition E' such that we know that we are E'-observers and such that we have empirical grounds for thinking that most E'-observers are not in simulations.)

For example, we can focus on our proximity to the dawn of the computer age, and use that as our E.

To be specific, let us focus on our relation to the date at which the first processor capable of operating at a clock speed of at least 1 MHz was created. Define a person's *computer age birth rank* as follows: The person whose birth was closest in time to the creation of the first such processor has rank 1; the person whose birth was second closest has rank 2; and so forth. For concreteness's sake, let us suppose that my computer age birth rank is 1 billion. Thus:

$E \equiv_{def}$ "My computer age birth rank is 1 billion."

It is plausible that any civilization that ends up running ancestor simulations at some point invents a processor with a clock speed of at least 1 MHz. It is also plausible that virtually every history in which such an invention occurs has a pre-posthuman epoch with at least 1 billion births; and hence, that these histories each contains some E-observer. Thus, condition (i) is satisfied. Furthermore, in all histories in which there is

some E-observer, there is exactly one E-observer, since at most one person can have a computer age birth rank of 1 billion. Thus, condition (ii) is satisfied.

Condition (iii) also appears to be satisfied. Although one can easily think of more specific centered propositions E' such that I know myself to be not only an E-observer but an E'-observer, this would enable the formulation of a defeater only if we had empirical grounds for thinking that most E'-observers are not in simulations. In fact, we are aware of no such grounds.[vi]

Conclusion

There is a technical glitch in the original presentation of the simulation argument. The glitch arises from the possibility that the average number of people living in the pre-posthuman phase might be different in civilizations that produce ancestor simulations than in civilizations that do not.

This glitch can be patched in at least two different ways, either of which secures the original conclusion. The first patch involves assuming that the average number of people living in the pre-posthuman phase is not *astronomically* greater for non-simulating civilizations than for civilizations that end up running significant numbers of ancestor-simulations. The second patch involves assuming that our type of experiences occur predominantly at a certain stage of history, so that even if the pre-posthuman phases lasted astronomically longer for non-simulating civilizations, they would nevertheless not on average contain vastly more people with our type of experiences than do the pre-posthuman phases of simulating civilizations.[vii]

Appendix

We illustrate how the first patch works. Assume that there have been only finitely many beings in the whole history of the universe. The number N is a given very large number such that there have been s civilizations that run at least N ancestor simulations each.[viii] The average number of pre-posthuman beings in them is H_s. There are n civilizations that did not run at least N ancestor simulations (because they run fewer, or because decided not to run any at all, or because they never reached a posthuman phase). The average number of pre-posthuman beings in them is H_n. Assume that:

$$\frac{H_n}{H_s} \leq \frac{N}{1000000}$$

Analysis:

1. We know that there have been exactly $nH_n + sH_s$ real beings. The number of simulated beings is unknown, but it is at least NsH_s.
2. We now estimate the fraction of beings that led simulated lives:

$$f_{sim} \geq \frac{NsH_s}{NsH_s + nH_n + sH_s}$$

$$= \frac{1}{1 + \frac{1}{N}\left(1 + \frac{nH_n}{sH_s}\right)}$$

$$\geq \frac{1}{1 + \frac{1}{N}\left(1 + \frac{n}{s}\frac{N}{1000000}\right)}$$

3. If $f_{sim} \geq 99\%$ then one of the statements of the simulation argument holds. Assume, then, that $f_{sim} < 99\%$ in order to see what follows.

$$\frac{1}{1 + \frac{1}{N}\left(1 + \frac{n}{s}\frac{N}{1000000}\right)} < \frac{99}{100}$$

$$\frac{100}{99} < 1 + \frac{1}{N}\left(1 + \frac{n}{s}\frac{N}{1000000}\right)$$

$$\frac{1}{99} - \frac{1}{N} < \frac{n}{s \cdot 1000000}$$

Given that $N > 9900$ we have

$$\frac{1}{99} - \frac{1}{N} > \frac{1}{99} - \frac{1}{9900} = \frac{1}{100}$$

and therefore

$$\frac{1}{100} < \frac{n}{s \cdot 1000000}$$

$$10000 \cdot s < n$$

This means that for every civilization that runs at least N simulations there are at least 10000 other which do not.

4. Let us write $n = a + b$, where a is the number of civilizations that never reached a posthuman phase, and b is the number of civilizations that did reach a posthuman phase but decided not to run simulations or to run simulations but fewer than N. If $b \geq 99 \cdot s$ then any posthuman civilization is no more than 1% likely to run a significant number of ancestor simulations, and the second statement of the simulation argument holds. Therefore all that remains to be checked is what happens when $b < 99 \cdot s$.

$$10000 \cdot s < a + b$$

$$10000 \cdot s < a + 99 \cdot s$$

$$9901 \cdot s < a$$

This means that for every civilization that runs a significant number of simulations there are more than 9900 civilizations that never reach the posthuman phase.

5. We now have the following estimates on the number of different types of civilizations:
    - s civilizations that reach posthumanity and run at least N simulations
    - no more than $99s$ civilizations that reach posthumanity but do not run simulations or run fewer than N
    - at least $9900s$ civilizations that never reach the posthuman phase

    The fraction of civilizations that never reach the posthuman phase is therefore at least

$$\frac{9900}{9900 + 99 + 1} = 99\%$$

    and the third statement of the simulation argument holds.

References

Barrow, J. D. "Living in a Simulated Universe" in *Universe or Multiverse*, ed. Bernard Carr (Cambridge University Press, 2007): 481-486

Brueckner, A. "The Simulation Argument Again", *Analysis*, Vol. 68 (2008): 224-226

Bostrom, N. "The Simulation Argument: A Reply to Weatherson", *Philosophical Quarterly*, Vol. 55 (2005): 90-97

Bostrom, N. "Are You Living in a Computer Simulation?" *Philosophical Quarterly*, Vol. 53 (2003): 243-255

Bostrom, N. "The Simulation Argument: Some Explanations", *Analysis*, Vol. 69 (2009): 458-461

Chalmers, D. "The Matrix as Metaphysics" in *Science Fiction and Philosophy*, ed. Susan Schneider (Wiley-Blackwell, 2009): 33-52

Hanson, R. "How to Live in a Simulation", *Journal of Evolution and Technology*, Vol. 7 (2001)

Jenkins, P. S. "Historical Simulations—Motivational, Ethical and Legal Issues", *Journal of Futures Studies*, Vol. 11 (2006): 23-42

Weatherson, B. "Are You a Sim?" *Philosophical Quarterly*, Vol. 53 (2003): 425-431

[i] "Posthuman stage" here refers loosely to a state in which technologies that we can already see are physically feasible have been developed, in particular powerful simulation technologies.

[ii] An ancestor simulation is a computer simulation that a posthuman civilization might run of its own history (and of variations thereof), in which brains are simulated with sufficient granularity to have conscious experiences. Throughout this article, by "computer simulations" we will mean "ancestor simulations". (It is also possible that we might live in a computer simulation that is not an ancestor simulation.)

[iii] For further details, see the original paper (Bostrom 2003).

[iv] E.g., (Hanson 2001, Weatherson 2003, Jenkins 2006, Barrow 2007, Brueckner 2008, Bostrom 2005, 2009, Chalmers 2009).

[v] We will assume throughout this paper that everything is finite, in order to avoid complications that arise when assigning probabilities and using indifference principles, such as the Self-Sampling Assumption, over infinite outcome spaces.

[vi] It could have been different. We could, for instance, have had some reason for thinking that all civilizations that ever create ancestor simulations do so within a year of creating their first 1 MHz processor—and that they terminate any of their simulations in which a (simulated) civilization has not yet attained the ability to create its own ancestor simulations within a few (simulated years) of it creating its first (simulated) 1 MHz processor. Had that been the case, then the fact that several decades have passed in our history since the creation of a 1 MHz processor without our civilization yet attaining the ability to create ancestor simulations would have constituted evidence against the simulation hypothesis. For we could then have formulated the defeater E': "My computer age birth rank is 1 billion and I am alive several decades after the creation of a 1 MHz processor in a civilization that has still not attained the ability to create ancestor simulations".

[vii] For a third way of patching the argument, one might try invoking the doomsday argument. Thus, one might argue that we have independent grounds for dismissing the hypothesis that there will be vastly many more people in our reference class in the future, since this would make our relatively early sequential position highly atypical. We do not propose this patch because we do not wish here to rely on the soundness of the doomsday argument. However, it is worth noting that if the doomsday argument were accepted, it could support the simulation argument, as follows: If the doomsday argument is used with the universal reference class, then it would support disjunct (1). If a more restrictive reference class is used that excludes posthumans, then it would support the claim that histories with vastly many more observers living in the pre-posthuman phase than have already been born in our history, are rare.

[viii] To get a sense for the magnitude of N: The original paper suggests that consideration of the theoretical limits of technology indicates that a technologically mature civilization, using the resources of a single planet, could create computational power sufficient for simulating the entire mental history of humankind by using less than a millionth of its processing power for only one second. Such a civilization might, of course, in principle last for many millions of years and might colonize many millions of planets.