The Simulation Argument Reconsidered

[This is a pre-copyedited, author-produced version of an article accepted for publication in Analysis *following peer review. The version of record is available online at: https://doi.org/10.1093/analys/anad048]*

1. Introduction

Some philosophers have begun to treat the hypothesis that we now exist within a simulation not merely as a skeptical possibility, but as somewhat probable[1] (Bostrom 2003, 2005a, 2005b, 2009; Bostrom & Kulczycki 2011; Chalmers 2022; Lewis 2013; White 2016). That this *simulation hypothesis* is somewhat probable—on the order of 20% (Bostrom 2005b)—has been defended most extensively by Nick Bostrom (2003; 2005a; 2005b; 2009). More recently, David Chalmers (2022: ch. 5) has suggested that there is a roughly 25% chance that we now inhabit a simulation. Notably, Bostrom (2003: 243) does not straightforwardly defend the conclusion that we inhabit a simulation, but rather the disjunctive conclusion that at least one of the following 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.

Bostrom argues that there is no basis for strongly favoring one of these alternatives and thus that it is fairly probable that we now inhabit a simulation—that we are sims. In this paper, I present

[1] Eric Schwitzgebel (2017), who regards this possibility as highly improbable, but worth taking into account for practical purposes, merits mention in this connection.

support for (2) and thereby argue that the simulation argument does not lend substantial credence to (3).

2. The simulation argument

We can now simulate environments containing vast numbers of agents engaging in complex interactions. Given projected advances in computing power, it is reasonable to expect that we will one day be able to create simulated agents that think and feel much as we do. One might doubt that sims will ever be able to feel, but the view that sims can be conscious has been defended by proponents of the simulation argument (Bostrom 2003; Chalmers 2022: ch. 5, ch. 15). Here, I assume that sims can be conscious[2]. In what follows, I will use "simulations" to refer narrowly to simulations involving such agents, unless otherwise stated. Additionally, following Bostrom (2003), I will use "simulations" to refer to "ancestor simulations"—simulations of posthuman civilizations' evolutionary histories, again unless otherwise stated. This focus is not due to the relevance of repeating details of history, but is meant rather to maintain focus on simulations whose scale and complexity resembles that of our universe.

Technological advances can be expected not only to allow for the creation of simulated worlds and persons, but the creation of simulated worlds and persons that vastly outnumber their non-simulated counterparts. I expect that there is no way to meaningfully estimate the number of simulations that able and willing posthuman civilizations would create, but even conservative estimates of this quantity yield the result that it is highly probable that we are sims inhabiting a simulated world. Thus, unless posthuman civilizations are either unable—either because such

[2] But for a skeptical perspective on this point, see Claus Beisbart (2014).

civilizations do not exist or lack the requisite technology—or unwilling to create simulated worlds with conscious inhabitants, it is probable that we are sims. In the absence of reason to think posthuman civilizations will either be unable or unwilling to create such worlds, proponents of the simulation argument recommend that the hypothesis that we are sims be assigned roughly equivalent probability to these alternatives.

3. Sims and sim-levels

Let us begin to assess the simulation argument by assuming that (1) and (2) are incorrect, and thus that the probability that we are now living in a simulation approaches 1. One might think, given these assumptions, that we are probably in a simulation created by a posthuman civilization that is not itself simulated. But this is overly simple. To see this, let us introduce some basic machinery.

Let us say that a simulation created directly by a non-simulated posthuman civilization is at sim-level$_1$ and that the non-simulated posthuman civilization is at sim-level$_0$. There is no obvious reason to suppose that sims cannot create further sims[3] (Bostrom 2003: 253). Indeed, discussion of the simulation argument often recognizes our progress toward the ability to create sims, while questioning whether we ourselves are sims. To recognize this possibility, let us say that a simulation created directly by simulated beings at sim-level$_1$ is at sim-level$_2$, sims created by sims at sim-level$_2$ are at sim-level$_3$, and so on. Supposing that we are probably sims, what sim-level do we most likely inhabit?

[3] One might object that sims cannot *really* do anything at all, including creating more sims (Brueckner 2008). However, it is unimportant for the purposes of this paper whether sim can create simulations or can merely simulate the creation of simulations. For simplicity, I write here as if sims can genuinely create simulations. See Bostrom (2009: 459-460) for a response to Brueckner's objection.

There is very little reason to suppose that we are at sim-level$_1$. After all, we are now supposing that the inhabitants of sim-level$_0$ are able and willing to create many simulations at sim-level$_1$. But thus far we have seen no reason to think the inhabitants at sim-level$_1$ will be any less able and willing to create simulations than the inhabitants of sim-level$_0$ are. Thus, we should expect the ratio of conscious beings at sim-level$_2$ to conscious beings at sim-level$_1$ to approximate the ratio of conscious beings at sim-level$_1$ to conscious beings at sim-level$_0$. The pattern can be expected to continue, such that simulations multiply in a branchlike fashion as we ascend to further levels of simulation. Assuming simulations multiply in this way, it is highly unlikely that we exist at some low sim-level.

It might be thought dubious that we are in a high-level simulation. If simulations multiply through levels as suggested above, we should also expect glitches to multiply in a similar fashion. Glitches may arise from avoidably faulty programming or from limitations in computing power. Let us start with the former. Inhabitants of sim-level$_0$ might, through avoidable mistakes, design environments at sim-level$_1$ that show signs of simulation. Any glitches at sim-level$_1$ need not, but may, cause further glitches at higher sim-levels, insofar as states of these are dependent on states of simulations at sim-level$_1$. Relatedly, as sim-levels increase, so too does the chance that glitches caused by limitations in computing power at sim-level$_0$ occur. As Bostrom recognizes, the computational cost for running simulations increases rapidly as further simulations are nested within them (2003: 253). Glitches are thus more probable at higher sim-levels than at lower sim-levels.

Here I will assume that no glitches in our apparent reality have been observed. Given this assumption, along with the proposition that glitches are more probable at higher sim-levels than lower ones, should we assume that we are not in a high-level simulation? No. Even if glitches

become increasingly common at higher sim-levels, this increase may be offset by the branching structure of simulations, such that most glitch-free simulations still occur at high sim-levels. Of course, the probabilities of programming errors and facts concerning the limitations of computing power may be such that, for some n, the number of glitch-free simulations at sim-level$_n$ is greater than the number of glitch-free simulations at sim-level$_{n+1}$. However, I doubt that we have anything resembling a reliable means of assessing how these probabilities and facts bear on the likelihood of glitches arising at various sim-levels. The only safe conclusion we can draw, I think, is that on the assumption that we are sims, we have no reason to believe ourselves to be at an especially low sim-level.

I have written thus far as if, if inhabitants of sim-level$_0$ can create simulations at sim-level$_1$, then inhabitants at and above sim-level$_1$ can create simulations at subsequent sim-levels. This is not to say that just any sim can create a simulation, only that sufficiently technologically sophisticated sims can do so. In short, the assumption is that if beings in base reality can create sims at sim-level$_1$ then, for any n>0, sims at sim-level$_n$ can, given technological sophistication comparable to that of the inhabitants of sim-level$_{n-1}$, create sims at sim-level$_{n+1}$. But this assumption is mistaken. As noted above, running a simulation at a given level will require vastly more computational power as further layers of simulation are nested within it. Thus, there is no guarantee that even extraordinarily technologically sophisticated sims can create further simulations (cf. Bostrom 2003: 253, 2009: 459). A sim might, by attempting to run a simulation, exceed the computational power of systems at lower sim-levels, or in base reality (Greene 2020; Tierney 2007). Indeed, assuming the computational resources devoted to simulations at the level of base reality are finite, there *must* be a limit to the layers of simulation these resources can

support. In the next section, I develop this point into an objection to the case for regarding (3) as probable.

4. The existential risks of simulation

The notion that we might someday create sophisticated simulations inhabited by conscious beings is a speculative possibility. Even if we could create such simulations, it is unclear how or if we could know that sims possess consciousness like our own. But let us set these difficulties aside and imagine a future—as far distant as one likes—in which we can create simulations that include sims that we know to possess consciousness similar to our own. In such a future, how should we regard (3)?

It is natural to suppose that the knowledge that we can create conscious sims should increase our suspicion that we ourselves are sims (Bostrom 2003: 253; Greene 2020: 494). But the discussion in section 3 complicates this picture. The attempt to run a simulation within a given sim-level might fail because of limits in computational power at that sim-level or because of limits in computational power at lower sim-levels, including in base reality. The attempt to run a simulation in base reality cannot fail due to insufficient computing power at lower sim-levels. For this reason, the knowledge that one is successfully running a simulation can be construed as evidence that one is not in a simulation.

This surprising conclusion should not be overstated. Although the computational costs of running nested simulations is high, nothing said here excludes the possibility that many layers of simulation might be layered over base reality. More to the point, nothing said here excludes the possibility that the individuals in our imagined future know that they have successfully run a

simulation, even though they are themselves sims. While the observation that a successful simulation has been created would undermine the view that one inhabits a simulation with constraints that prevent the creation of further simulations, it would not undermine the view that one inhabits a simulation without such constraints[4]. Thus, the observation that a complex simulation has successfully been created is evidence for one version of the simulation hypothesis and against another. Without independent reasons to favor one version over the other, this hypothetical observation should be regarded as evidentially equivocal with respect to a generic form of the simulation hypothesis. I will soon offer some reason to think that, if we inhabit a simulation, we inhabit a simulation that has constraints against further simulations, and thus some reason to treat the hypothetical observation as evidence against the simulation hypothesis.

Still, consideration of this imagined future does little by itself to indicate how we, who have not successfully run simulations populated with conscious beings, should regard (3). Even if the creation of such a simulation constituted strong evidence against (3), this is not evidence we have. However, consideration of sim-levels points to an argument for (2) and in this way undermines the case for (3).

As we have seen, the attempt to execute a simulation may fail due to a lack of computational power at the sim-level within which the simulation is attempted or at a lower sim-level. For example, the attempt to run a simulation at sim-level$_3$ may fail because the systems responsible for directly simulating sim-level$_1$—that is, computers in base reality—or systems in the intervening levels are not sufficiently powerful to support so many nested simulations. Such a failure would likely have limited consequences for individuals in base reality, but might have

[4] Thanks to an anonymous referee for pressing me on this point.

disastrous consequences for all those sims that depend for their existence on the continuation of simulations running in base reality. Preston Greene recognizes this point, noting that "ancestor simulations entail a termination risk to those that create them" (2020: 493). For this reason, there exist strong reasons of self-preservation for any beings who do not know themselves to be in base reality not to simulate complex worlds inhabited by conscious beings. Assuming we do not know ourselves to be in base reality, we can reasonably take this ignorance to be shared.

The upshot of the preceding argument is that posthuman civilizations are unlikely to create simulations involving conscious beings, on the grounds that doing so will be perceived as an existential risk (Greene 2020). Consequently, there is reason to doubt that we ourselves are simulated. This argument might seem far-fetched. Unless we assume that the generation of simulations is extraordinarily costly—an assumption for which there is no clear basis—we ought to allow that constraints on computational resources do not eliminate the possibility of multiple layers of simulation. Why, then, would it be reasonable for any particular posthuman civilization to fear that its creation of a simulation will be the back-breaking straw (henceforth a 'back-breaking simulation')? Three responses are in order. First, if this reasoning is sound for any posthuman civilization, it will be sound for *every* posthuman civilization. But, given finite constraints on computational resources, some civilizations really are at risk. Second, the termination risk to a given posthuman civilization does not depend on that civilization *directly* attempting to create the back-breaking simulation. For example, the attempt to create sim-level$_3$ by sim-level$_2$ might put sim-level$_1$ at risk. Thus, even if one is justified in assigning a low probability to the proposition that creating simulations is directly risky, one cannot safely assume that creating simulations is not *indirectly* risky, insofar as it enables the creation of a back-breaking simulation at a higher sim-level. Third, and relatedly, termination of a simulation might occur not

because the actual limits on computational resources are reached, but instead because those at some lower level of simulation recognize risks to themselves, and therefore terminate any simulation attempting to create further simulations (cf. Greene 2020: 494). It is for this reason that, if we have reason to think we are in a simulation, we have reason to think we are in a simulation with constraints against the creation of further simulations. More generally, any would-be simulators have much to fear not only from resource constraints on simulations, but also from the defensive actions of beings in base reality and lower levels of simulation, if there are any such beings.

Greene (2020: 496-499) makes much of the relationship between would-be simulators and any beings that might exist at lower levels of simulation. In particular, Greene emphasizes how the bidirectional dependency between the inclination to create sims and the probability that we ourselves are sims generates a decision problem in which the expected utility of creating simulations is unstable. Whereas Greene focuses on what a posthuman civilizations' inclinations might tell it about the inclinations of civilizations that might exist at *lower* levels of simulation, it is also important to recognize what a civilizations' inclinations would suggest about the inclinations of those at *higher* levels of simulation. Insofar as we are inclined to create simulations, we should expect at least some sims we create to be inclined to create further simulations, and so on[5]. However, as we have seen, the creation of layers of simulation *cannot* be carried out *ad infinitum*. One should thus expect that if one generates a simulation, a back-breaking simulation will eventually be created—at least so long as other would-be simulators are left to their own devices. Given the risk of termination either due to computational limits or the defensive actions

[5] Admittedly, simulators might reduce their sims' inclinations to create sims by including, within simulations they create, evidence of simulation. However, any such evidence would alter the behavior of the sims, thereby reducing the value of the simulation (cf. Greene 2020: 495, 505).

of those in lower sim-levels, posthuman civilizations always have a powerful incentive to avoid creating simulations. Consequently, there is reason to regard (2) as highly probable and thus to reject the case for assigning a high probability to (3). In the terminology of section 2, posthuman civilizations are likely to be unwilling to create complex simulations.

5. Objections and replies

I now consider three objections to the preceding arguments. First, one might argue that, if the preceding argument is reason to doubt that posthuman civilizations would create complex simulations involving conscious beings, it is also reason to doubt that such civilizations would create relatively simple simulations. The latter conclusion being implausible, one might thus doubt the former[6]. However, there are two key differences between simulations that involve conscious beings and those that do not. First, even a single simulation of the former type can be expected to be vastly more computationally expensive than a simulation of the latter type. Second, as I have emphasized above, the creation of one layer of conscious sims opens the door to the creation of further layers (cf. Greene 2020: 503)—a process that *cannot* be repeated indefinitely given finite computational resources. According to a second objection, it is one thing to say that there is a strong practical reason not to simulate complex worlds, and another to say that posthuman civilizations are therefore unlikely to do so. It might be thought that such civilizations cannot safely be expected to avoid risks to their own survival. Evidence for such pessimism might be gleaned from human recklessness toward the existential dangers of climate change and nuclear arms. However, such pessimism should not be overstated. First, apparent human apathy toward climate

[6] Thanks to an anonymous referee for raising this point.

change and nuclear weapons is plausibly due more to ignorance and the difficulty of collective action than to genuine disregard for existential threats**.** In particular, whereas climate and nuclear weapons policies can plausibly be modeled as prisoner's dilemmas—in that each state actor is rational to pursue strategies that, take together, lead to suboptimal outcomes for all—individual parties have compelling reasons to unilaterally avoid creating complex simulations. Although there might be strategic reasons to create complex simulations—for intelligence purposes, for example—these benefits are at least not obviously sufficient to outweigh the consideration that one's own creation of a complex simulation entails a termination risk. Moreover, there is no obvious reason why any intelligence advantages to be gained by complicated simulations could not be achieved through less computationally costly means, including simulations with shorter histories and without conscious beings. Additionally, it seems likely that those with the ability to run simulations would be best acquainted with the risks of doing so. Finally, complex simulations will be run, if at all, by advanced civilizations, whose very survival is some evidence of their responsiveness to existential risks. Thus, it is reasonable to conclude that the dangers of running complex simulations would be appreciated by those civilizations positioned to do so. Even if would-be simulators are unmoved by ethical considerations (cf. Bostrom 2003: 252; Greene 2020: 502), they are likely to be moved by reasons of self-preservation.

While we now expect nested simulations to be exceptionally computationally expensive, future technological advances might in principle significantly alter the perception of these costs[7]. This could occur if, for example, we come to identify less computationally intense methods for supporting simulations within simulations or if we discover that the resources at our disposal are greater than initially thought. Two responses are in order. First, any discoveries we might make

[7] Thanks to an anonymous referee for raising this objection.

will effectively be discoveries concerning the methods and resources available within our perceived reality, which we cannot with full confidence regard as non-simulated. Thus, even if we pursue simulation strategies that appear non-costly relative to the resources available, this appearance may be illusory with respect to the resources available in base reality or intermediate simulations. So long as we take seriously the possibility that we are simulated, our empirical discoveries cannot assure us of the safety of generating further simulations. If we are indeed simulated, any such discoveries will pertain at least in the first instance to *simulated* computational resources, and need not correspond to the actual resources devoted to our own simulation. Second, as we have seen, even if technologies available in base reality can support may layers of simulation, they cannot support *infinite* layers. This point—coupled with the point emphasized above that, if a posthuman civilization creates posthuman sims, it should expect this pattern of creation to continue—implies that reductions in the cost of simulation relative to available resources do not significantly reduce the termination risk.

6. Concluding remarks

I have argued that there is good reason to accept (2) in the disjunctive conclusion of Bostrom's (2003) simulation argument. In doing so, I have offered reason to reject the assignment of a high probability to the claim that we are sims. The thrust of this argument is that any beings that are both concerned for their own survival and unsure whether they are sims have good reason not to run complex simulations. It does not follow that we are not sims, only that the assignment of a relatively high probability to our being sims is premature.

To conclude this essay, I want to draw attention to an irony in the case for (2). It is the inability to rule out that one is in a simulation that will lead one to regard running complex simulations as a potential existential risk. In this way, the inability to rule out the simulation hypothesis, and the recognition of this shared predicament, contributes to the case against assignment of a high probability to that hypothesis[8].

References

Beisbart, C. (2014). Are we sims? How computer simulations represent and what this means for the simulation argument. *The Monist*, 97(3), 399-417.

Bostrom, N. (2003). Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211): 243-255.

Bostrom, N. (2005a). The simulation argument: Reply to Brian Weatherson. *Philosophical Quarterly*. 55: 90-97.

Bostrom, N. (2005b). Why make a Matrix? And why you might be in one. In W. Irwin (Ed.) *More Matrix and Philosophy*, pp. 81–92. New York: Open Court.

Bostrom, N. (2009). The simulation argument: Some explanations. *Analysis*, 69(3), 458-461.

Bostrom, N. & Kulczycki, M. (2011). A patch for the simulation argument. *Analysis*, 71(1), 54-61.

[8] I would like to thank two anonymous referees for their careful and extensive feedback on previous drafts of this paper.

Brueckner, A. (2008). The simulation argument again. *Analysis*, 68(3), 224-226.

Chalmers, D. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: W.W. Norton & Company.

Greene, P. (2020). The termination risks of simulation science. *Erkenntnis*, 85, 489-509.

Lewis, P.J. (2013). The doomsday argument and the simulation argument. *Synthese*, 190, 4009-4022.

Schwitzgebel, E. (2017). 1% Skepticism. *Noûs*, 51(2), 271-290.

Tierney, J. (2007). Our lives, controlled from some guy's couch. *New York Times*, 14.8.2007. https://www.nytimes.com/2007/08/14/science/14tier.html

White, J. (2016). Simulation, self-extinction, and philosophy in the service of human civilization. *AI & Society*, 31(2), 171-190.