Simulation expectation

Teruji Thomas (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2021

GPI Working Paper No . 16-2021

Simulation Expectation

Teruji Thomas*

Abstract

I present a new argument that we are much more likely to be living in a computer simulation than in the ground-level of reality. (Similar arguments can be marshalled for the view that we are more likely to be Boltzmann brains than ordinary people, but I focus on the case of simulations.) I explain how this argument overcomes some objections to Bostrom's classic argument for the same conclusion. I also consider to what extent the argument depends upon an internalist conception of evidence, and I refute the common line of thought that finding many simulations being run—or running them ourselves—must increase the odds that we are in a simulation.

Keywords: simulation hypothesis, anthropic reasoning, termination risk, Boltzmann brains, self-location.

1 Introduction

Here's a way the world might be. At some point there exist conscious beings whose experience of the world is much like ours—let's just call them *people*. And at some point in their history, these people run computer simulations of whole worlds, so powerful that these worlds are inhabited by other such conscious beings—let's call them *simulant people*. And these simulant people might even run further simulations on their (simulant) computers, containing other simulant people, and so on. Only if we live

*GPI working paper version May 2023. Your comments and questions welcome at teruji.thomas@philosophy.ox.ac.uk. In addition to my colleagues at GPI, I'm grateful to Joe Carlsmith, Kenny Easwaran, Ben Garfinkel, Maria Lasonen-Aarnio, Harvey Lederman, and Jeff Russell for helpful discussions, and to Natasha Oughton, Elliott Thornley, John Mori, and Riley Harris for assistance.

in the non-simulant, ground-level of reality (if there even is such a thing!) are we ourselves non-simulant people.[1]

I will present an argument for

> **SIM.** It is much more likely that I am a simulant person than a non-simulant person.

Bostrom (2003) presents a closely related argument (with a correction in Bostrom and Kulczycki (2011)), known as the Simulation Argument. It has inspired a great deal of philosophical and popular discussion. Strictly speaking, the Simulation Argument, as Bostrom construes it, is *not* an argument for SIM, but it does *suggest* such an argument, which has been the focus for most discussion. I'll henceforth call this suggested argument 'the classic argument'—remembering that Bostrom himself does not endorse it. Arguments similar to the classic one have been given for other hypotheses, for example the hypothesis that I am more likely to be a Boltzmann brain than an ordinary person[2]—but I'll focus on simulations here.

I'll give a reconstruction of the classic argument in section 2, but I can already explain why I find it unsatisfactory. The argument depends on the premiss that my current evidence entails (or at least strongly supports) the following empirical claim:

> **HIGH RATIO.** The ratio of simulant to non-simulant people is high.

Bostrom did not argue for HIGH RATIO, and he appears to end up with roughly a 1/3 credence in it. For all the classic argument says, this is compatible with a 1/3 credence that I am a simulant person and a 2/3 credence that I am a non-simulant person. More generally, for all the classic argument says, it could be much less likely that I'm a simulant person than a non-simulant person, as long as HIGH RATIO is unlikely. Thus, one problem for arguing for SIM from HIGH RATIO is that we don't have much reason to be confident in HIGH RATIO. Perhaps more interestingly, it is hard

[1]Without claiming it's a settled question, I'll follow the literature in assuming that there might be simulant people with mental lives relevantly like our own.

[2]See Carroll (2020); Kotzen (2020); Dogramaci (2020) for recent discussion. In this paper, I'm happy *not* to count Boltzmann brains and other 'freak observers' (Crawford, 2013) as 'people'. Indeed, I'll sometimes assume that, if I'm a non-simulant person, then my experiences are generally verdical, and the world is basically as we ordinarily take it to be. On the other hand, this means that (i) it may not be certain on my evidence that I am a person at all; (ii) SIM does not say anything about whether I am more likely to be a simulant person or a non-simulant freak observer—a question I simply set aside.

to see how we *could* have evidence that strongly supports HIGH RATIO, in a way that is compatible with SIM. For HIGH RATIO is, in part, a claim about the number of people in the ground-level of reality, and, if we ourselves are in a simulation, that's not the sort of thing about which we plausibly have much evidence.[3]

So my own argument will not be based on HIGH RATIO, and this is the main way in which it improves upon the classic argument. Instead, the analogous premiss in my argument is

> **HIGH EXPECTATION.** *Conditional* on my being a non-simulant person, the *expected* ratio of simulant to non-simulant people *in the appropriate reference class* is high.

I'll say more about what 'in the appropriate reference class' means later, but the basic idea is to consider only people who inhabit worlds in broad strokes like our own: they live on minor variants of 21st century Earth.[4]

Although I will not argue for HIGH EXPECTATION in any detail, I believe it is *this* claim, and not HIGH RATIO, that is supported by the empirical considerations adduced in Bostrom's paper, and especially by his claims about feasible computing power. Here's the idea. Suppose I'm a non-simulant person. It may be quite unlikely that our descendants will run simulations of their ancestral 21st century. But (Bostrom argues) they could in principle run *enormously* many, at negligible cost to themselves, and even on a whim.[5] Those simulated 21st century people would be in the reference class. So there's at least a small probability that the ratio of simulant to non-simulant people in the reference class is enormous. As long as the probability is not *too* small, HIGH EXPECTATION is true. In contrast, this line of reasoning does not particularly support HIGH RATIO.

[3]Versions of this objection are pressed by Birch (2013), especially his section 3, and Crawford (2013). See also Dogramaci (2020) for similar objections in the context of Boltzmann brains. One can thus read this paper as a response to theirs: I give an argument for SIM on roughly the same grounds but immune to this problem. Brueckner (2008) objects more simply that the probability of HIGH RATIO is inscrutable; this does not affect my argument either.

[4]Perhaps HIGH RATIO should also involve a reference class restriction of some sort, e.g. to beings with what Bostrom calls 'human-type experiences'; I've just assumed that this is baked into my vague characterisation of 'people'.

[5]According to Bostrom (2003, 247–8), 'A single [planetary-mass] computer could simulate the entire mental history of humankind…by using less than one millionth of its processing power for one second. A posthuman civilization may eventually build an astronomical number of such computers.'

The main advance in this paper is to replace HIGH RATIO by HIGH EXPECTATION, thus giving us more reason to take seriously the possibility that we are simulant people. I develop the argument in sections 2 to 3. This move does not (however) resolve some other issues which I will discuss in section 4. I will especially consider, and tentatively respond to, a worry raised by Weatherson (2003) about what evidence I can have if I happen to be in the ground-level of reality. And I will refute the common line of thought that finding lots of simulations being run must increase the odds that I myself am in one.

Let me conclude this introduction by explaining why I think assessing SIM is an important project, even from a practical point of view. Jenkins (2006, 23) puts the point in a usefully provocative way. Given the classic argument, he writes, it is

> highly probable that we are a form of artificial intelligence inhabiting one of these simulations. To avoid stacking (i.e. simulations within simulations), the termination of these simulations is likely to be the point in history when the technology to create them first became widely available.... Long range planning beyond this date would therefore be futile.

While this line of thought is highly speculative, it does seem right that living in a simulation may carry with it some distinctive risks (e.g., the risk of simulation termination) as well some distinctive opportunities (e.g., to mitigate termination risk by restricting technological progress).[6] These are, in other words, speculations that may be worth taking seriously if there really is a good argument for SIM. My goal here is to give the best and clearest argument I can.

2 Preliminaries

In this section, I'll explain the framework I'll be using in this paper, and then I'll indicate the general form of the argument to follow—including a brief reconstruction of the classic argument for SIM.

[6] For more recent and careful work exploring these and similar worries, see Tomasik (2016) and especially Greene (2020). See Chalmers (2022) for a wider-ranging exploration of life in virtual worlds.

2.1 The Framework

First, the basic framework. Questions of likelihood could be interpreted in different ways. I will use a Bayesian framework, in which the natural question is what is supported, probabilistically, by my total evidence. So, SIM is the claim that my current evidence strongly supports the hypothesis that I'm a simulant person over the hypothesis that I'm a non-simulant person.

My evidence in this sense is the sort of thing that appears in Bayesian norms like conditionalization: a proposition. I will represent propositions in the form *I am F*—or, for brevity in formulas, ιF—to emphasise that they can have an indexical or 'self-locating' aspect. We can think of F itself as a property (cf. Lewis, 1979). For example, if F is the property of being a simulant person, then ιF is the proposition that I am a simulant person.

Instead of just tracking my current credences, I assume that the facts (or, if you prefer, my subjective judgments) about evidential support are encoded by a probability measure (or Popper function) Pr, my *ur prior*. The probability that I am F, given my total evidence that I am E, is thus the conditional probability $\Pr(\iota F \mid \iota E)$.

That said, it is sometimes more convenient to talk in terms of odds rather than probabilities. (The specific reason it is convenient will appear at the end of §3.3.) The odds that I am F rather than G, given that I am E, are defined to be

$$\mathrm{Odds}(\iota F/\iota G \mid \iota E) \stackrel{\text{def}}{=} \frac{\Pr(\iota F \mid \iota E)}{\Pr(\iota G \mid \iota E)}.$$

So SIM says that the odds are high that I'm a simulant person (ιF) rather than a non-simulant person (ιG), given my current evidence (ιE).[7]

2.2 The Classic Argument

With this framework in place, the general style of my argument—and of the classic argument—can be anticipated from the following uncontroversial, if informal, piece of reasoning.

> I'm a smoker, and one out of fifteen smokers develops lung cancer. So, if I had no other information, I would say that

[7] As mentioned in footnote 4, I don't assume that it's certain that I'm a person; thus *I'm not a simulant person*, $\neg\iota F$, may not be equivalent to *I'm a non-simulant person*, ιG. If one doesn't care about this issue, one can replace G by not-F in all the relevant places.

> the probability that *I'll* develop lung cancer is 1/15. I do have other information, but none of it is particularly relevant to whether I'll develop lung cancer. So, taking all my evidence into account, the probability is still about 1/15.

This reasoning depends on (1) the identification of a 'reference class'—the class of smokers; (2) an empirical premiss about frequencies within that reference class—the frequency of lung cancer; (3) a premiss about what could be concluded if one had no other information; (4) the premiss that, although one does have other information, it is mostly irrelevant to the question at hand.

The classic argument follows this logic quite closely.[8] (1) For the reference class, we can take the class of all people (defined in whatever way makes the argument most compelling). (2) The empirical premiss is HIGH RATIO, to the effect that the vast majority of this reference class are simulants. (3) If I had no other information, this would make it extremely likely that, if I am a person at all, then I am a simulant person. (4) Finally (Bostrom suggests) though I do have other evidence, none of it tells very strongly *against* the thesis that I am a simulant person rather than a non-simulant person. Therefore, even taking all my evidence into account, it is still far more likely that I am a simulant person than a non-simulant person.

In the introduction, I raised doubts about premiss (2). I also think that premiss (4), while somewhat plausible, is difficult to assess. I will discuss some worries about it in section 4, focusing (however) on the analogous step in my own argument for SIM, to which I now turn.

3 The Argument

The structure of the argument will parallel those described in section 2.

3.1 The Reference Class

As I mentioned in the introduction, I take the reference class to consist of people living on minor variants of 21st century Earth. At least, this is a provisional definition that is enough to get the argument going. We can

[8] The form of the argument reconstructed here is more overt in the exchange between Weatherson (2003) and Bostrom (2005) than in the original paper (Bostrom, 2003).

amend it in whatever way makes the premisses most compelling—a point on which I'll elaborate as we go, especially in section 4.

3.2 The Empirical Premiss

The first premiss of the argument is HIGH EXPECTATION, which is basically an empirical claim about the ratio of simulant to non-simulant people in the reference class. (More exactly, it is a claim about my evidence about this ratio.)

Let me give a more formal statement of the premiss. Let F be the property of being a simulant person, G the property of being a non-simulant person, and R the property of being in the reference class. (For short I will sometimes refer to R itself as 'the reference class'.) Let $\mathrm{Rat}_R^{F/G}$ denote the ratio of Fs to Gs among all Rs (i.e., here, the ratio of simulant to non-simulant people in the reference class). Then the expected ratio of simulant to non-simulant people in the reference class, conditional on my being a non-simulant person, is defined by the sum

$$\mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G) \stackrel{\text{def}}{=} \sum_r r \times \Pr(\mathrm{Rat}_R^{F/G} = r \mid \iota E \,\&\, \iota G).$$

Here r ranges over the countably many candidate values for $\mathrm{Rat}_R^{F/G}$, i.e., all non-negative rational numbers.[9] With this definition, the premiss is

HIGH EXPECTATION. $\mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G)$ is high.

I already sketched in the introduction why HIGH EXPECTATION might be true with respect to the specified reference class. Of course, 'high' is vague and context-dependent, but this doesn't matter, since the argument will conclude (in equation 1 below) that the odds I'm a simulant person are also high in whatever sense is on the table.

3.3 Frequency-Based Reasoning

Suppose that (strangely enough) my only evidence is that I'm a smoker and that that 1 in 15 smokers develop lung cancer. How likely is it, on this evidence, that *I* develop lung cancer? The probability is 1/15, calibrated

[9] A non-trivial assumption I'm making is that the number of Rs is finite. Then $\mathrm{Rat}_R^{F/G}$ is also finite, on the condition that I myself am both R and G. In universes with infinite populations, frequency-based reasoning faces general and serious problems (Arntzenius and Dorr, 2017).

to the frequency of cancer-developers among all smokers. Or, in other words, the odds are 1/14, calibrated to the ratio of cancer-developers to non-developers among all smokers.

Here's the obvious generalization, formulated for any properties F, G, and R, and any non-negative rational number r.

CALIBRATION. $\text{Odds}(\iota F/\iota G \mid \iota R \,\&\, \text{Rat}_R^{F/G} = r) = r$.

The second premiss of my argument is that CALIBRATION holds for all r, when F is the property of being a simulant person, G is the property of being a non-simulant person, and R is the reference class.[10]

In the appendix (and in the example below) I prove that CALIBRATION has the following non-obvious formal consequence, which is what's directly relevant to my argument:

CALIBRATION*. $\text{Odds}(\iota F/\iota G \mid \iota R) \geq \mathbb{E}(\text{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G)$.

So, if my total evidence consisted in the facts that characterise the reference class (ιR), then the odds that I am a simulant person rather than a non-simulant person would be at least as great as the expected ratio of simulant to non-simulant people, conditional on my being a non-simulant person.

By the way, CALIBRATION* explains why I generally focus on ratios and odds. The principle is more cumbersome when expressed in terms of frequencies and probabilities.

3.4 Admissibility

The final premiss claims that I do not have much relevant evidence beyond the fact that I am in the reference class. We can include in the proposition ιR facts about the laws of physics, the limits of computational power, human psychology, and the apparent trajectory of civilization (all, that is, for

[10] Here, I will rest on the intuitive plausibility of CALIBRATION; my argument for SIM does not seem much worse than the classic argument in this respect. However, CALIBRATION is a weak principle of the sort often used in the context of anthropic reasoning and self-locating credences; see Manley (2014) for an overview, and Bostrom (2002) for discussion of the best-known such principle in the simulation-adjacent literature, the (generally much stronger) 'self-sampling assumption'.

Note that I do not endorse CALIBRATION for *every* triple of properties. For example, as I discuss in Thomas (2021), one can come up with counterexamples if F, G, and R are not presented in what Chalmers (2004) calls 'neutral' terms. In that paper I defend a more elaborate principle, which arguably reduces to CALIBRATION in the case at hand.

the world I inhabit, not necessarily for the ground-level of reality). And I simply don't have much to go on, beyond these general facts, when it comes to assessing (i) whether I am a simulant person or not, or (ii) the ratio of simulant to non-simulant people among the *R*s. The more specific details of my life, like my name and address, certainly don't seem relevant.

At first, let us suppose that my additional evidence is *completely* irrelevant in these two ways. Then, formally, the reference class *R* satisfies the following two conditions (in which case I'll say that *R* is *admissible*).

ADMISSIBILITY.

(i) $\text{Odds}(\iota F/\iota G \mid \iota E) = \text{Odds}(\iota F/\iota G \mid \iota R)$;

(ii) $\mathbb{E}(\text{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G) = \mathbb{E}(\text{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G)$.

The point of ADMISSIBILITY is that it allows us transform each side of CALIBRATION* to get:

$$\text{Odds}(\iota F/\iota G \mid \iota E) \geq \mathbb{E}(\text{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G). \tag{1}$$

This is like CALIBRATION*, except that it takes my total evidence (ιE) into account. At this point we can appeal to HIGH EXPECTATION. It tells us that the expected ratio is high; so, therefore, are the odds.

Now, more plausibly, conditions (i) and (ii) will only be approximately true, so that *R* is only approximately admissible. But this does not make much difference. Indeed, as long as (i) $\text{Odds}(\iota F/\iota G|\iota E)$ is not much smaller than $\text{Odds}(\iota F/\iota G \mid \iota R)$, and (ii) $\mathbb{E}(\text{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G)$ is not much larger than $\mathbb{E}(\text{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G)$, it follows from CALIBRATION* that $\text{Odds}(\iota F/\iota G \mid \iota E)$ is not much smaller than $\mathbb{E}(\text{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G)$. If the latter is high, so must be the former.

This concludes the argument for SIM, except for the proof of CALIBRATION*, which I give in the appendix. In the rest of this section, I'll give a 'proof by example'—that is, a toy example that illustrates the logic of the proof, which many readers may find more illuminating than the derivation in full generality.

Example

Suppose that the entirety of my evidence is that I am *R*—it is this situation that CALIBRATION* concerns. To get a simple example, let me stipulate the following details. There are two possibilities, *A* and *B*, compatible with the assumption that I am both *R* and a non-simulant person. Either

(A) There are no simulant people who are Rs, so $\text{Rat}_R^{F/G} = 0$; or else

(B) There are 100 simulant people for every non-simulant person among the Rs, so $\text{Rat}_R^{F/G} = 100$.

Suppose more specifically that

($*$) Conditional on my being a non-simulant person, A has probability 0.9 and B has probability 0.1.

It follows that the expected ratio of simulant to non-simulant people is

$$\begin{aligned}\mathbb{E}(\text{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G) &= 0 \times \Pr(A \mid \iota R \,\&\, \iota G) + 100 \times \Pr(B \mid \iota R \,\&\, \iota G) \\ &= 0 \times 0.9 + 100 \times 0.1 \\ &= 10.\end{aligned}$$

On the other hand, we have to calculate the odds that I'm a simulant person. Let's start by comparing the probabilities of the following three propositions, conditional on my total evidence, ιR.

($A \,\&\, \iota G$) A is true and I'm a non-simulant person.

($B \,\&\, \iota G$) B is true and I'm a non-simulant person.

($B \,\&\, \iota F$) B is true and I'm a simulant person.

First, $A \,\&\, \iota G$ is nine times more likely than $B \,\&\, \iota G$. How come? We can use the formal identity $\text{Odds}(A \,\&\, \iota G / B \,\&\, \iota G \mid \iota R) = \text{Odds}(A/B \mid \iota R \,\&\, \iota G)$. The latter equals 9, according to ($*$).

Second, $B \,\&\, \iota F$ is 100 times more likely than $B \,\&\, \iota G$. Here we can use the formal identity $\text{Odds}(B \,\&\, \iota F / B \,\&\, \iota G \mid \iota R) = \text{Odds}(\iota F/\iota G \mid \iota R \,\&\, B)$. The latter equals 100, by CALIBRATION.

It follows from the previous two claims that $B \,\&\, \iota F$ is 10 times more likely than $A \,\&\, \iota G$ and $B \,\&\, \iota G$ put together. But $A \,\&\, \iota G$ and $B \,\&\, \iota G$ exhaust the possibilities for my being a non-simulant person, while $B \,\&\, \iota F$ does not necessarily exhaust the possibilities for my being a simulant person. So it's *at least* 10 times more likely than I'm a simulant person than a non-simulant person. We have established

$$\text{Odds}(\iota F/\iota G \mid \iota R) \geq 10 = \mathbb{E}(\text{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G).$$

Thus we have derived the key claim, CALIBRATION*, in this special case.

4 Discussion

The main subtlety in my argument for SIM concerns the choice of the reference class *R*. I have so far used a rough characterization of *R* in terms of 'minor variants of 21st century Earth', which is enough to show how such an argument can get off the ground (whereas the classic argument seems hopeless). However, since the conclusion, SIM, does not depend on *R*, we can choose *R* in any way that makes the premisses compelling. Of these premisses, I will not discuss CALIBRATION further, since it reflects a very general form of frequency-based reasoning that should work for most reasonable ways of specifying *R*. But there is an interesting tension between the remaining premisses, HIGH EXPECTATION and ADMISSIBILITY, which I will consider in this section. To be clear: essentially the same tension exists in the classic argument for SIM, and I don't think my own argument is much, if at all, worse off in this respect. But the following points are still important for understanding the limitations of the argument I've proposed.

The basic tension is as follows. We can guarantee that ADMISSIBILITY holds by packing a lot of information into *R*, making the proposition ιR close to my total evidence; but this makes HIGH EXPECTATION less plausible. For example, ADMISSIBILITY holds trivially if we take $R = E$, for then I have no further evidence at all beyond ιR. But, supposing that I'm a non-simulant person, and even if there are a vast number of simulant worlds, it's unlikely that there are many, or perhaps any, simulant people with the very specific property *E*. So the expected ratio will be relatively low. On the other hand, if we take *R* to be much weaker than my total evidence—applying, perhaps, to most people who have ever lived—then HIGH EXPECTATION may be arguable, but ADMISSIBILITY becomes less plausible, or simply hard to assess. I will study two examples of this phenomenon.

Before proceding, let me make explicit a point that could easily get lost in the high-level conceptual discussion: the actual numbers matter! If the expected ratio in HIGH EXPECTATION is extremely high, then the odds I'm a simulant will also be high, unless ADMISSIBILITY dramatically fails.

4.1 The Limits of Appearances

Let's suppose that I am in fact a non-simulant person, living in the ground-level of reality. And for concreteness let's consider the view that my evidence consists of what I know (Williamson, 1997). One thing I know—or so we would ordinarily say—is that the world around me is more than 10 billion years old, and that throughout many of those billions of years, vast expanses of the universe were lifeless. In constrast, we might think that most simulated worlds are *not* vast and mostly lifeless (VML) in this way: it would be easier to simulate a relatively small, inhabited portion of the universe and make it *appear* that the world was VML.

This leads to a dilemma: should we include in the reference class property *R* the fact that the world is VML? The first horn: if we do include it, then the expected ratio of simulant to non-simulant *R*s will be relatively low, since most simulant worlds are not VML. The second horn: if we don't include it, then the first clause of ADMISSIBILITY will fail, since I have additional evidence (the fact that my world is VML) that confirms I'm a non-simulant.[11]

I see this dilemma as the philosophically most interesting objection to my argument, so let me take some time to sketch how it might be defused. One could, of course, try to argue on empirical grounds that HIGH EXPECTATION is true even if we include the claim that the world is VML, and similar claims, in *R*. But otherwise I see three promising, if closely related, conceptual strategies.

The first is to adopt an internalist conception of evidence, according to which my evidence is essentially just a matter of how things appear to me. So, even if I'm a non-simulant person, my evidence does not entail that my world is VML; perhaps it only entails that my world *appears* VML. There is no problem with including this fact about appearances in the reference class, insofar as it would not greatly reduce the expected ratio.[12]

---

[11] The claim that my world is VML is, of course, only an example. Weatherson (2003, 429–430) raises a similar objection in the context of the classic argument for SIM, though he focuses on evidence about more proximate matters like the fact that I have real eyeballs. Chalmers (2005) and Bostrom (2005) provide some ammunition to respond to Weatherson: they argue that, if I am in a sufficiently rich simulation, then I *do* have real eyeballs and the like; in general, being in such a simulation is not the sort of radically skeptical scenario in which my most commonplace beliefs turn out to be false. I focus on the claim that my world is VML because, as Chalmers acknowledges in the various scenarios considered in his section 8, the world beyond my immediate macroscopic environs is something about which I, as a simulant, could more easily be mistaken.

[12] Weatherson (2003, 430) suggests that a version of the objection applies to some internalists as well. To place some limits on the discussion, I have to set aside the interesting question of what the best internalist views, and other externalist views, would say here; I don't want to suggest that they go scot free. In particular, it is important to remember that the basic Bayesian framework requires one's evidence to be a proposition, and on this score it is not always clear what internalists have in mind.

Alternatively, we could retain an externalist conception of evidence (as I myself would prefer), and even retain the view that my evidence is my knowledge, but argue that there are special features of the case that prevent me from knowing that my world is VML. Bostrom (2005, 94) suggests a view along these lines, in the context of the classic argument for SIM. In a possible world where there is a high ratio of simulant to non-simulant people, 'illusions are ubiquitous': 'almost all people…have perceptions which, if interpreted naïvely, are misleading about [certain] facts'. And if we know that's how things are, then we cannot trust the appearances: knowledge of HIGH RATIO (or maybe the mere truth of HIGH RATIO) prevents me from knowing that my world is vast and mostly lifeless.

Be that as it may, it is initially unclear whether Bostrom's move helps defend *my* version of the argument, the whole point of which is not to rely on HIGH RATIO. I'm not considering a situation in which I *know* that the ratio of simulant to non-simulant people in my reference class is high, or even necessarily one in which *it's true* that this ratio is high. It's just that the *expected* ratio is high, conditional on my being a non-simulant person. To escape the dilemma in Bostrom's way, I would need to argue that in *this* epistemic situation I cannot know that my environment is VML.

While I lack a compelling argument for this position, I think it is defensible. Let me explain. Suppose we include in the specification of the reference class only claims about how the world appears, such as the claim that the world appears to be VML. The dilemma primarily concerns the first clause of ADMISSIBILITY, not the second, which seems relatively safe.[13] Using only that second clause, CALIBRATION* still leads to the inequality

$$\mathrm{Odds}(\iota F/\iota G \mid \iota R) \geq \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G).$$

Continuing to assume that the expected ratio on the right-hand side is high, this inequality shows that the general facts about how things appear (those involved in the fact that I am R) strongly support the view that I am a simulant rather than a non-simulant person. Let us also grant to the objector that no simulated worlds are really VML. Then, going by these general facts about how things appear—*including* the fact that my world appears to be VML—it's highly unlikely that my world is VML. This does

[13] That second clause involves the condition that I am a non-simulant person, and we may assume that the appearances are generally veridical on this condition (cf. footnote 4). So, as far as the second clause goes, there is not much space between the claim that my world appears to be VML and the claim that it actually is.

seem like the type of situation that could prevent my knowing that my world is VML.[14]

The third strategy for avoiding the dilemma is a variation on the second. We can make a partial retreat. There are different questions to be asked: questions about what I'm in a position to know and what that knowledge would support; but also questions about what credences a reasonable but fallible subject would have or ought to have.[15] We might concede to the objector that, *if* in fact I'm a non-simulant person, then I'm in a position to know that my world is VML, and my evidence would then support high confidence that I'm a non-simulant person. But, in another sense, I'm still in an unusually bad position to affirm the antecedent: as I've just argued, the facts about how things broadly appear support its negation. This may make it reasonable to adopt credences that are out of line with what's supported by my evidence, or—perhaps better—to abandon my belief, and hence my evidence-as-knowledge, that the world is, in the relevant ways, exactly as it seems.

4.2 The Limits of Future Evidence

While I sketched an argument for HIGH EXPECTATION in §1, the success of that argument depends on empirical factors that I have not examined closely. On the other hand, even if that argument fails, it seems plausible that we could gain evidence in the future that would make HIGH EXPECTATION true. However, contrary to some claims in the literature, the overall effect of such evidence can be difficult to gauge, and would not necessarily tend to confirm that we are simulant people.

For example, what would happen if I were to find a secret lab running quadrillions of whole-world simulations, or if I were to find myself about to run such simulations? Bostrom (2006, p. 9) and Greene (2020) both claim that this should make me confident that I am a simulant person.[16]

---

[14]To be more explicit: what's highly unlikely is the conditional that, if I'm a person at all, then my world is VML. It's knowledge of this conditional that is required to support the claim that I'm a non-simulant person as opposed to a simulant person.

[15]Without meaning to attribute to them any position on the current topic, I have in mind the kinds of distinctions drawn by Lasonen Aarnio (2010) and Schoenfield (2012).

[16]As Greene discusses at length, this claim could be practically relevant to whether we *should* run simulations, at least given evidential decision theory. Bostrom and Greene are both thinking about the classic argument, and so for them the idea is that finding the secret lab would make me confident in HIGH RATIO. That's hardly obvious: again, if I'm in a simulation, finding a secret lab would not tell me much about the number of non-

In terms of my own argument, I am happy to concede that such evidence could make HIGH EXPECTATION true. The problem is that the reference class may no longer be admissible.

Before explaining that problem in detail, I think it is worth taking a step back. Instead of applying a general argument for SIM, we can just ask what kind of an update the new evidence would provide, using updating norms like conditionalization. The answer in part depends on what we take the new evidence to be. On the externalist account, perhaps I get to know that there *really are* many simulations being run within my environment. But as Lewis (2013) explains, this evidence *could* quite powerfully confirm that I myself am non-simulant, since it's plausible on combinatorial grounds that the vast majority of simulations do not themselves contain many simulations.[17] Even if we focus on how things appear, the situation is unclear. Perhaps the new evidence is only that there *appear* to be many simulations within my environment. (After all, the lab computers bleep and bloop, but I don't have direct access to what's going on inside.) In order to strongly confirm that I'm a simulant person, this appearance must be much more likely on the supposition that I'm a simulant than on the supposition that I'm a non-simulant, against the background of my current evidence.[18] I don't see why it would be.

---

simulant people. But even granting this claim, step (4) in the classic argument might well fail with respect to my new evidence, for reasons similar to those I discuss below.

[17]To illustrate, suppose that people in the ground-level world created 100 simulated worlds, each of which contained a further 100 simulated worlds (and the process stopped there). Then 99% of worlds would contain no simulations. Curiously, Lewis seems to construe the key point of the Simulation Argument as being a probabilistic inference from *there are no simulations in my world* to *I'm in a simulation*.

[18]Here I'm appealing to the fact that when my evidence strengthens from ιE to $\iota E'$, the odds that I am F rather than G get multiplied by the *Bayes factor*

$$\text{Bayes}_{\iota F/\iota G}(\iota E' \mid \iota E) = \frac{\Pr(\iota E' \mid \iota F \,\&\, \iota E)}{\Pr(\iota E' \mid \iota G \,\&\, \iota E)}.$$

In words: the odds increase insofar as it is more likely that I am E' on the supposition that I am F than on the supposition that I am G, against the background evidence that I am E.

It may be worth reiterating the point made by Crawford (2013, 262) that finding lots of (real or apparent) simulations would hardly make it likely that I'm living in one of *those* simulations. The situation may become complicated if the discovered simulations involve phenomenal duplicates of me, leaving me in the position of 'Dr. Evil' (Elga, 2004). Alternatively, perhaps there could be metaphysical cycles of simulations within simulations.

Returning to the admissibility constraint, there are two ways to look at the situation. If we keep the reference class fixed, then my new evidence may well make the expected ratio high. But insofar as this new evidence supports the claim that I am a simulant person, the first clause of ADMISSIBILITY must fail. On the other hand, we could analyse things in terms of a new reference class. We must choose the reference class property *R* so that the fact that I am *R* includes most of my relevant evidence. Obviously, one relevant part of my evidence is that there are (or there appear to be) many simulations run within my world. So we have to include something to that effect within the definition of *R*—we have to consider a more limited reference class. At that point, the problem is that the expected ratio with respect to this new reference class need not be higher than before.

5 Conclusion

The situation is this. Bostrom's Simulation Argument suggests a further argument, based on HIGH RATIO, that we ourselves are simulants. But it's hard to see why we should be confident of HIGH RATIO, and it's hard to see how we could get good evidence about the ratio of simulant people to non-simulant people if we ourselves are in a simulation. This paper indicates how to get around both these points. First, it suffices to have evidence about the ratio *on the condition* that we are non-simulant people. Second, we need not be confident, on that condition, that the ratio is high; it suffices that the *expected* ratio is high. Because of this, the resulting argument for SIM is much more troubling (if that is the right word) than the one based on HIGH RATIO.

That argument is the main point of this paper. However, I have also indicated some ways to resist it. A key step in the argument is to assume that there is some reference class property *R* that is both roughly admissible and that leads to a high expected ratio. One could resist this on strictly empirical grounds (for example, though not only, by rejecting Bostrom's claims about the limits of computational power or the uses to which our descendents seem likely to put it). Or perhaps one could resist this step on more conceptual grounds by adopting an appropriate theory of evidence. Tentatively, though, I have suggested that giving high odds that we are simulant people may still be reasonable *even if* we are non-simulant people and our evidence supports the opposite conclusion. Finally, I have argued that finding a large number of simulations being run, or otherwise acquiring evidence that raises the expected ratio, would not necessarily raise the odds that we are simulant beings.

Appendix: Proof of CALIBRATION*

The main technical step in the argument for SIM is the derivation of

CALIBRATION*. $\mathrm{Odds}(\iota F/\iota G \mid \iota R) \geq \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G)$.

This only requires CALIBRATION. I illustrated the derivation in section 3 with a toy example, but here I consider the general case.

Start from the observation that

$$\Pr(\iota F \mid \iota R) \geq \sum_r \Pr(\iota F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \iota R) \tag{2}$$

where the sum is over all non-negative rational r. I will consider each term in this sum, one at a time. Let $a(r)$ denote the rth term:

$$a(r) \stackrel{\text{def}}{=} \Pr(\iota F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \iota R).$$

Recall the definition of conditional probability:

$$\Pr(X \mid Y) = \Pr(X \,\&\, Y)/\Pr(Y).$$

I will use this to repackage conditional probabilities without detailed explanation. (If one prefers to treat conditional probabilities as primitive, one could instead use the axioms for Popper functions.) As a first application, for each candidate value of r we have

$$a(r) = \Pr(\iota F \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) \times \Pr(\mathrm{Rat}_R^{F/G} = r \mid \iota R).$$

Now we can express the first factor on the right-hand side in terms of odds and then apply CALIBRATION:

$$\begin{aligned} \Pr(\iota F \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) &= \mathrm{Odds}(\iota F/\iota G \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ &\qquad \times \Pr(\iota G \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ &= r \times \Pr(\iota G \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r). \end{aligned}$$

Substituting this into the previous equation, we have

$$a(r) = r \times \Pr(\iota G \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) \times \Pr(\mathrm{Rat}_R^{F/G} = r \mid \iota R).$$

Reuse the definition of conditional probability to repackage this:

$$a(r) = r \times \Pr(\mathrm{Rat}_R^{F/G} = r \mid \iota R \,\&\, \iota G) \times \Pr(\iota G \mid \iota R).$$

If, as in the inequality (2), we sum $a(r)$ over all values of r, we get

$$\sum_r a(r) = \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G) \times \Pr(\iota G \mid \iota R).$$

Substituting this into (2), we get

$$\Pr(\iota F \mid \iota R) \geq \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota R \,\&\, \iota G) \times \Pr(\iota G \mid \iota R).$$

Dividing through by $\Pr(\iota G \mid \iota R)$ yields CALIBRATION*.

References

Arntzenius, F. and Dorr, C. (2017). Self-locating priors and cosmological measures. In Chamcham, K., Barrow, J., Saunders, S., and Silk, J., editors, *The Philosophy of Cosmology*, pages 396–428. Cambridge: Cambridge University Press.

Birch, J. (2013). On the 'simulation argument' and selective scepticism. *Erkenntnis*, 78(1):95–107.

Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

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

Bostrom, N. (2005). The simulation argument: Reply to Weatherson. *The Philosophical Quarterly*, 55(218):90–97.

Bostrom, N. (2006). Do we live in a computer simulation? *NewScientist*, 192(2579).

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

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

Carroll, S. M. (2020). Why Boltzmann brains are bad. In Dasgupta, S., Weslake, B., and Dotan, R., editors, *Current Controversies in Philosophy of Science*. Routledge.

Chalmers, D. J. (2004). Epistemic two-dimensional semantics. *Philosophical Studies*, 118(1-2):153–226.

Chalmers, D. J. (2005). The matrix as metaphysics. In Grau, C., editor, *Philosophers Explore the Matrix*, pages 132–176. Oxford University Press.

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. Norton.

Crawford, L. (2013). Freak observers and the simulation argument. *Ratio*, 26(3):250–264.

Dogramaci, S. (2020). Does my total evidence support that I'm a Boltzmann brain? *Philosophical Studies*, 177(12):3717–3723.

Elga, A. (2004). Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*, 69(2):383–396.

Greene, P. (2020). The termination risks of simulation science. *Erkenntnis*, 85(2):489–509.

Jenkins, P. (2006). Historical simulations-motivational, ethical and legal issues. *Journal of Futures Studies*, 11(1):23–42.

Kotzen, M. (2020). What follows from the possibility of Boltzmann brains? In Dasgupta, S., Weslake, B., and Dotan, R., editors, *Current Controversies in Philosophy of Science*. Routledge.

Lasonen Aarnio, M. (2010). Unreasonable knowledge. *Philosophical Perspectives*, 24(1):1–21.

Lewis, D. (1979). Attitudes *de dicto* and *de se*. *Philosophical Review*, 88(4):513–543.

Lewis, P. J. (2013). The doomsday argument and the simulation argument. *Synthese*, 190(18):4009–4022.

Manley, D. (2014). On being a random sample. Manuscript available on Manley's website.

Schoenfield, M. (2012). Chilling out on epistemic rationality: A defense of imprecise credences. *Philosophical Studies*, 158(2):197–219.

Thomas, T. (2021). Doomsday and objective chance. Working paper.

Tomasik, B. (2016). How the simulation argument dampens future fanaticism.

Weatherson, B. (2003). Are you a sim? *The Philosophical Quarterly*, 53(212):425–431.

Williamson, T. (1997). Knowledge as evidence. *Mind*, 106(424):1–25.