Simulation expectation

Teruji Thomas (Global Priorities Institute, University of Oxford)

Global Priorities Institute | September 2021

GPI Working Paper No . 16-2021

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Simulation Expectation

Teruji Thomas*

Abstract

I present a new argument for the claim that I'm much more likely to be a person living in a computer simulation than a person living in the ground-level of reality. I consider whether this argument can be blocked by an externalist view of what my evidence supports, and I urge caution against the easy assumption that actually finding lots of simulations would increase the odds that I myself am in one.

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 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

*Working paper version September 2021. Your comments and questions welcome at teru.thomas@oxon.org. In addition to my colleagues at GPI, I'm grateful to Maria Lasonen-Aarnio, Kenny Easwaran, Harvey Lederman, Jeff Russell, Ben Garfinkel, and Elliott Thornley for helpful discussions.

[1]Without claiming that it's a settled question, I'll just assume that there might be simulant people with mental lives relevantly like our own. In general, I'll leave it loose what counts as a 'person'. But I won't assume that it's certain on my evidence that I am a person at all. That allows us to exclude 'freak observers' like Boltzmann brains, and to set aside the question of whether I am most likely overall to be a non-simulant freak observer, for discussion of which see Crawford (2013). Relatedly, I will assume that, if I am a non-simulant person, then my experiences are generally veridical.

> **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. However, the Simulation Argument is not an argument for SIM. The relevant part of Bostrom's argument, slightly reconstructed, simply claims

> **CONDITIONAL SIM.** Conditional on the ratio of simulant people to non-simulant people being high, it is much more likely that I am a simulant person than a non-simulant person.

The ratio here involves people who exist at any point in time, not just at present. One way to argue for SIM would be to add to the Simulation Argument an argument for the condition in CONDITIONAL SIM. We could, in other words, argue for

> **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 Simulation 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 Simulation 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 to see how we *could* reasonably be confident in 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 could have much evidence.[2]

So I will not argue for HIGH RATIO. Instead, the analogous premiss in my argument is

[2]This is the main idea of Birch (2013), especially his section 3, and related to the second objection of Crawford (2013). 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. For further skeptical worries, e.g. associated with the possibility that I'm a Boltzmann brain, see footnote 1.

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

The restriction to 'my reference class' is a delicate one, to be discussed later, but the rough idea is to consider only people to whom the world appears in broad strokes like our own: they live on minor variants of 21st century Earth.[3]

Although I will not argue for HIGH EXPECTATION in any detail, it seems fairly plausible, if we grant Bostrom's 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 they could in principle run *enormously* many, at negligible cost to themselves, and even on a whim.[4] Those simulated 21st century people would be in my reference class. So there's at least a small probability that the ratio of simulant to non-simulant people in my 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.

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 5. This move does not (however) resolve some other issues which I will discuss in section 6. I will especially consider, and tentatively respond to, a worry raised by Weatherson (2003) about the nature of my evidence if I am in fact a non-simulant person. And I will urge caution against the easy assumption that actually finding lots of simulations being run would increase the odds that I am in a simulation.

2 The 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. And my evidence in this sense is

[3] CONDITIONAL SIM may 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'.

[4] 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 sort of thing that appears in Bayesian norms like conditionalization: a proposition. (More probing questions about the nature of my evidence will become relevant in section 6.)

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 as a property, a point of view made popular by Lewis (1979). 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 Cr, the *ur prior*. The probability that I am F, given my total evidence that I am E, is thus the conditional probability $\mathrm{Cr}(\iota F \mid \iota E)$.

However, it is sometimes more convenient to talk in terms of odds rather than probabilities. The odds that I am F rather than G, given that I am E, are defined to be

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

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).[5]

Just as probabilities are connected to frequencies, the odds that I am F rather than G, given that I am E, are connected to the *ratio* between the number of instances of F and the number of instances of G among all instances of E. (I'll say more about what the connection is in the next section.) Now, I'm interested in the situation across all time, not just at the present, and a single thing can be F at some times and not at others. So it's natural to count 'instances' in a time-weighted way: if something is F for five units of time, consecutively or not, then that counts as five instances of F. The ratio of interest, which I'll denote by $\mathrm{Rat}_E^{F/G}$, is then the ratio between the number of instances of $F \,\&\, E$ and the number of instances of $G \,\&\, E$, counted in this time-weighted way. But it is briefer and usually harmless to refer to $\mathrm{Rat}_E^{F/G}$ as the ratio of Fs to Gs among the Es.[6]

[5] As noted in footnote 1, 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.

[6] As I emphasise in Thomas (2021), there are some further complications that arise from hyperintensional distinctions. In this paper I just assume that we can regiment the discussion so that such subtleties don't arise. In brief, though, for each possible world w, one really wants to get at the ratio between (i) the number of epistemic scenarios in which w is actual in the primary intension of 'I am $F \& E$' and (ii) the analogous number for 'I am $G \& E$'. Another issue I am ignoring is how to measure time within a simulation (or even in the real

3 Warm Up: Frequency-Based Reasoning

To motivate the basic moves in my argument, I will first consider a much simpler case, analogous to the situation in which one is certain of HIGH RATIO.

> I am in a cohort of 100 patients. Among us, some have blood type *A* and some have blood type *B*. What are the odds that I have type *A* rather than *B*, conditional on the hypothesis that the ratio of patients with type-*A* blood to patients with type-*B* blood is *r*?

The obvious answer (though not necessarily the correct one) is *r*: the ratio of frequencies within the 'reference class' of patients. Formally, if we let *F* be the property of having type-*A* blood, *G* be the property of having type-*B* blood, *R* the property of being one of the 100 patients, and ιE my total evidence, then

$$\mathrm{Odds}(\iota F/\iota G \mid \iota E \,\&\, \mathrm{Rat}_R^{F/G} = r) = r. \qquad (*)$$

To work towards this answer we can start from a general principle of indifference.[7]

> **INDIFFERENCE.** For any properties *F*, *G*, and *R*, and any number *r*,
>
> $$\mathrm{Odds}(\iota F/\iota G \mid \iota R \,\&\, \mathrm{Rat}_R^{F/G} = r) = r.$$

For example, given *only* that I am one of the 100 patients and that, among us, the ratio of frequencies is 4-to-1, the odds that I have type-*A* rather than type-*B* blood are also 4-to-1. However, INDIFFERENCE does not quite cover the original example, because there I may have additional background evidence. And, indeed, whether *r* is the correct answer depends on this evidence. For example, if I know that blood type is heritable and that my mother had type-*B* blood, and I don't know anything similar about the other patients, then the odds that I have type *A* rather than *B* could well be less

world, given relativistic physics). Presumably the right way tracks each person's subjective experience of time, and thus the apparent passage of time *within* each simulation.

[7] Since formulations in terms of odds and ratios are somewhat uncommon, it may be worth noticing that INDIFFERENCE is equivalent to a similar principle about probabilities and frequencies: given only that I am *R* and that the frequency of *F*s among *R*s is *p*, the probability that I am *F* is *p*. While this can (indeed) be called a principle of indifference, in Thomas (2021) I explain how such principles can be motivated by a certain view of objective chance.

than r. But if my background evidence does not tell me anything important about myself in contrast to the other patients, then the obvious answer would seem to be correct. Finding out the frequency of blood types among the patients would 'screen off', or render irrelevant, whatever other evidence I might have as to my own blood type.

To formulate the relevant condition, it's convenient to recall that, when my evidence strengthens from ιE to $\iota E'$, the odds that I am F rather than G change by the *Bayes factor*

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

That is to say,

$$\text{Odds}(\iota F/\iota G \mid \iota E') = \text{Bayes}_{\iota F/\iota G}(\iota E' \mid \iota E) \times \text{Odds}(\iota F/\iota G \mid \iota E).$$

As this indicates, $\text{Bayes}_{\iota F/\iota G}(\iota E' \mid \iota E)$ is a natural measure of how strongly the new evidence that I am E' would confirm or disconfirm that I am F rather than G, given my background evidence that I am E. When it equals 1, the new evidence neither confirms nor disconfirms. By definition, though, the Bayes factor is just a matter of whether it is more likely that I am E' given that I am F than given that I am G, against the background evidence that I am E.

Formally, then, the issue is just whether the reference class of patients is *admissible* in the following sense:

> **ADMISSIBILITY.** R is admissible (with respect to E, F, G, and r) if and only if E entails R and
>
> $$\text{Bayes}_{\iota F/\iota G}(\iota E \mid \iota R \,\&\, \text{Rat}_R^{F/G} = r) = 1.$$

This makes precise the idea that I have no evidence as to whether I am F rather than G that would not be screened off by my evidence that I am R and the information that the ratio of Fs to Gs among the Rs is r. Together, INDIFFERENCE and ADMISSIBILITY entail the desired claim $(*)$.

The argument for SIM from HIGH RATIO follows this same basic kind of reasoning. Indeed, CONDITIONAL SIM is a loose version of $(*)$, where F is the property of being a simulant person, G is the property of being a non-simulant person, and R is the property of being a person. (It is 'loose' in that it refers only to *high* ratios rather than to any particular ratio.) The basic motivation for CONDITIONAL SIM is the thought that this R is at least approximately admissible for a relevant range of high values of r.[8]

[8]This is essentially the motivation given in Bostrom (2005).

4 The Abstract Result

My own argument, though more complicated, follows the same strategy as the one in the previous section. We should identify a reference class property R, entailed by my total evidence, such that (i) R is admissible, or close enough; and (ii) we have usable information about the ratio of Fs to Gs among the Rs. In this section, I'll present the main result in abstract form, returning in section 5 to the crucial issue of how we should choose R.

I will actually give two versions of the result. First, I will state a version that uses the exact premisses of the previous section, INDIFFERENCE and ADMISSIBILITY. However, I prefer a second version that replaces ADMISSIBILITY by a more manageable condition, ADMISSIBILITY*.

Remember that HIGH EXPECTATION concerns an expected ratio. If again F is the property of being a simulant person, G the property of being a non-simulant person, and R the property of being in my reference class, then the relevant expectation is defined by the sum

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

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

> **Main Result, Version 1.** Assume that INDIFFERENCE is true. If R is admissible with respect to E, F, G, and every r, then
>
> $$\mathrm{Odds}(\iota F/\iota G \mid \iota E) \geq \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G).$$
>
> In particular, if the expected ratio is high, so are the odds.

Now, ADMISSIBILITY is a very strict condition: it imposes an equation for every value of r. But for the inference from 'high expected ratio' to 'high odds', we only need R to be *approximately* admissible, for then the stated inequality will not be badly wrong. Moreover, we only need R to be approximately admissible for *most* values of r. Since it is not easy to make precise what sort of approximation is allowable in these terms, I think it is more insightful to invoke an alternative admissibility condition that does not quantify over r.[10]

[9] A non-trivial assumption I'm making is that the number of Es is finite. Then, if I myself am E and G, $\mathrm{Rat}_E^{F/G}$ must be finite. In universes with infinite populations, frequency-based reasoning faces general and serious problems (Arntzenius and Dorr, 2017).

[10] Note however that ADMISSIBILITY* does not follow from ADMISSIBILITY, nor vice versa.

ADMISSIBILITY*. *R* is admissible* (with respect to *E*, *F*, and *G*) if and only if *E* entails *R* and

(i) $\mathrm{Odds}(\iota F/\iota G \mid \iota E) = \mathrm{Odds}(\iota F/\iota G \mid \iota R)$; in other words,

$$\mathrm{Bayes}_{\iota F/\iota G}(\iota E \mid \iota R) = 1.$$

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

These conditions say that my evidence relevant to whether I am *F* rather than *G*, and my evidence relevant to the expected ratio, are exhausted by my evidence that I am *R*.

We still have:

Main Result, Version 2. Assume that INDIFFERENCE is true. If *R* is admissible* with respect to *E*, *F*, and *G*, then

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

In particular, if the expected ratio is high, so are the odds.

Again, for the inference from 'high expected ratio' to 'high odds', we only need ADMISSIBILITY* to hold up to a reasonable approximation. The details are given along with the derivation of the main results in the appendix. For now, I'll just note the basic strategy for proving this version of the main result. From INDIFFERENCE alone we can obtain:

INDIFFERENCE*. For any properties *F*, *G*, and *R*,

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

The inequality here is essentially the same as in the main result, but with *R* instead of *E*. The main result itself follows if we apply the first clause of ADMISSIBILITY* to the left-hand side of the inequality and the second clause to the right.

5 Applying the Main Result

Let us take *F* to be the property of being a simulant person, *G* the property of being a non-simulant person, and ιE my total evidence. The main result says that, for any admissible* reference class, HIGH EXPECTATION implies SIM. The remaining difficulty is to specify the reference-class property *R*.

Whatever E and F may be, there is always *some* admissible* property: if nothing else, we can take R to be the same as E.[11] For each such R, the main result gives a lower bound on the odds that I am a simulant person. Any lower bound is somewhat interesting, but the pressing issue is whether we can find an R that is at least approximately admissible*, and such that the lower bound (i.e. the expected ratio appearing in HIGH EXPECTATION) is high.

In the introduction, I suggested that we could choose R so that the people who are R are the people living in parallel versions of 21st century Earth. I then sketched an argument that the expected ratio is high. Can we argue that this R, spelled out in some way, is at least approximately admissible*? This is a subtle issue, with more discussion to come in the next section, but here is the basic argument. We can specify the proposition that I am R as a conjunction of facts about how the world appears to be going: it can include the apparent facts about the laws of physics and the limits of computational power, about human psychology, the broad strokes of history, and the trajectory of civilization. And I simply don't have much to go on, beyond these broad appearances, when it comes to assessing whether I am a simulant person or not, and when it comes to assessing the ratio of simulant to non-simulant people among the Rs. (The more specific details of my life, like my name and address, certainly don't seem relevant.) But that's just to say that both clauses of ADMISSIBILITY* hold to a good approximation.

This completes my argument for SIM.

6 Discussion

The argument just completed remains loose in some obvious respects. But I hope I have shown how such an argument can get off the ground, whereas, I suggested in the introduction, the argument for SIM from HIGH RATIO seems hopeless. While there is much more that could be done to evaluate and potentially bolster this argument, I will focus on two issues which—while not new, and not unique to *my* argument for SIM—strike me as especially interesting and important. They also usefully illustrate the delicate role of the reference class R.

The first issue has to do with the argument for ADMISSIBILITY*, and the second with the argument for HIGH EXPECTATION. Before proceding, however, let me make explicit a point that could easily get lost in the high-level

[11] This is true for ADMISSIBILITY as well as ADMISSIBILITY*. I'll focus on the second condition, but it would also be interesting to consider the first.

conceptual discussion: these two conditions interact, and 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.

6.1 The Limits of Appearances

When arguing in §5 for the ADMISSIBILITY* of a suitably specified *R*, I relied on the idea that my relevant evidence is substantially a matter of how the world appears to be going; these appearances are already entailed by the fact that I am *R*. But as Weatherson (2003) argues, in the context of CONDITIONAL SIM, the success of this move may depend on difficult questions about the nature of my evidence. Those aren't questions I'm going to settle in this paper; instead, I would prefer to piggyback on the replies to Weatherson made available by Bostrom (2005) and Chalmers (2005). However, on balance, I must admit that the available replies are not as strong as I might wish, so let me indicate how things play out with respect to the arguments of this paper, and offer an updated, if tentative, response.

For concreteness, I'll focus on the simple externalist view that my evidence is the conjunction of what I know. (I don't thereby mean to suggest that other views of evidence are either off the table or off the hook.) Here's the problem. Assume that I am in fact a non-simulant person. One thing I know is that the world around me is billions of years old, and that throughout many of those billions of years, vast expanses of the universe were lifeless. But it would be surprising if someone created such a spatiotemporally vast and mostly lifeless world as a simulation—why not just create the good bits, and make it *appear* that the world is vast? So while indeed it *appears* to simulant *R*s that they live in such a world, my evidence—that I *do* live in such a world—is more likely given that I'm a non-simulant person and an *R* than it is given that I'm a simulant person and an *R*. This and other considerations of a similar kind could add up to mean that the Bayes factor appearing in the first clause of ADMISSIBILITY* is much lower than 1, so that we cannot apply the main result.[12] And, on the other hand, redefining *R* to include (among other things) the fact that my world is vast and mostly lifeless could significantly reduce the expected ratio of simulant to non-simulant people among the *R*s.

[12] A *low* Bayes factor, i.e. less than 1, means that my evidence beyond the fact that I am *R* *disconfirms* that I'm a simulant person instead than a non-simulant person. As explained in the appendix, the inference from 'high expectation' to 'high odds' works if the Bayes factor (there denoted by α) is not very low compared to other parameters.

Of course, we could respond to this worry by rejecting the view that evidence is knowledge, and arguing that some favoured alternative avoids any similar problem. However, Bostrom and Chalmers suggest two more conciliatory responses, which would be preferable if they succeed.

First, a more radical form of the objection is that I know I have hands—real hands!—but a simulant does not have real hands. So, while indeed it *appears* to simulant *R*s that they have hands, this evidence—my really having hands—supports the claim that I am a non-simulant person by entirely ruling out the possibility that I am a simulant person.[13] Chalmers gives us ammunition to respond to this form of the objection, with Bostrom (2005, 94–5) concurring: according to him, I *do* have real hands if I am in a sufficiently rich simulation. 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.

However, conceding all this to Chalmers, it does not completely see off the objection, even in its radical form. For even if appropriately rich simulations are possible in principle, they may not be typical; then, my having hands would still tend to confirm that I am a non-simulant person. This seems especially relevant when we consider our purported knowledge of less proximate matters, beyond our immediate and present environs. Again, it's possible *in principle* that we're in a simulation so rich and extensive that our beliefs about the vast lifelessness of our spacetime come out true; but the objection, as I originally posed it, trades only on the thought that such simulations are not the norm. Indeed, Chalmers acknowledges (in the various scenarios considered in his section 8) that the world beyond my immediate macroscopic environs is something about which I, as a simulant, could easily be misled.

Bostrom (2005, 94) suggests a different response, which may be more promising. 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 unclear whether this move helps defend *my* version

[13] This is the surface reading of Weatherson's objection. He also gives an 'internalist' version of the objection based on the claim that simulant people don't have real eyeballs. To place some limits on the discussion, I have to set aside the interesting question of how the story goes for other specific views of my evidence; it particular, I don't want to suggest than internalists go scot free.

of the argument, the whole point of which was 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 objection in Bostrom's way, I would need to argue that in *this* epistemic situation I cannot know that my environment is vast and mostly lifeless.

While I lack a compelling argument for this position, I think it is defensible. Let me explain. Remember that the second clause of ADMISSIBILITY* concerns what is expected on the condition that I am a non-simulant person. On this condition, we can assume that the appearances are generally veridical (cf. footnote 1), including, to continue the example, the appearance that the world is vast and mostly lifeless (henceforth 'VML'). This is why I have construed Weatherson's objection as targeting the first clause of ADMISSIBILITY*, but not the second. Now, using only that second clause, the argument for the main result sketched at the end of section 4 still yields the inequality

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

This differs from the main result in that the odds on the left-hand side are conditional on $\imath R$ rather than on my total evidence $\imath E$. 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 seem like the type of situation that could prevent my knowing that my world is VML.[14]

We could also strengthen this defense by making 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 Weatherson's 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

[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 Aarnio (2010) and Schoenfield (2012).

my evidence would then support high confidence that I'm a non-simulant person. But, in another sense, I'm still in an unsually 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.

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

Step back for a moment. Instead of applying a general argument for SIM we can just ask what kind of an update the new evidence would provide: what is the Bayes factor? This 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.

[16] As Greene discusses at length, this claim could be practically relevant to whether we *should* run simulations, at least given evidential decision theory.

[17] Just by way of illustration, 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*.

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 purportedly going on inside.) Ask again: what's the Bayes factor? Is this appearance much more likely given that I'm a simulant person than given that I'm a non-simulant person, against the background of my current evidence? It's hard to see why.[18]

Returning to the admissibility constraint, there are two ways to look at the situation. If I 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 such that the relevant part of my new total evidence does not go far beyond the fact that I am R. 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. But the expected ratio with respect to this new reference class need not be higher than before.

7 Conclusion

The situation is this. Bostrom's Simulation Argument suggests a further argument, based on HIGH RATIO, that we ourselves are simulants. But (i) it's hard to see why we should be confident of HIGH RATIO, and (ii) 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 simulant. 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 admissi-

[18] It may be worth reiterating the point made by Crawford (2013, 262) that finding lots of (real or apparent) simulations could 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)).

ble* 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 (*pace* Chalmers and Bostrom) perhaps one could resist it 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 most reasonable even if, assuming that we are non-simulant people, our evidence in principle supports the opposite conclusion. Finally, I have argued that acquiring new evidence, like finding a large number of simulations being run, would not necessarily further confirm the view that we are in a simulation.

Appendix: Proof of the Main Result

The main result was stated in two versions in section 4: one based on ADMISSIBILITY and the other on ADMISSIBILITY*. I will develop these results in a way that makes clearer what sorts of approximations are allowable.

First, define λ to be the largest number such that:

$$\text{For all } r,\ \text{Bayes}_{\iota F/\iota G}(\iota E \mid \iota R \,\&\, \text{Rat}_R^{F/G} = r) \geq \lambda.$$

ADMISSIBILITY implies that $\lambda = 1$. A generalization of the main result (version 1) is thus:

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

In particular, if the expected ratio is high, and none of the Bayes factors is too small, then the odds are high.

Second, define α and β by

$$\alpha = \frac{\text{Odds}(\iota F/\iota G \mid \iota E)}{\text{Odds}(\iota F/\iota G \mid \iota R)} = \text{Bayes}_{\iota F/\iota G}(\iota E \mid \iota R)$$

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

ADMISSIBILITY* says that $\alpha = \beta = 1$. Note that β, while not a Bayes factor, is still a measure of how strongly my evidence bears on $\text{Rat}_R^{F/G}$, given that I am R. A generalization of the main result (version 2) is

$$\text{Odds}(\iota F/\iota G \mid \iota E) \geq \frac{\alpha}{\beta} \times \mathbb{E}(\text{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G). \tag{2}$$

So, if the expected ratio is high and β is not much bigger than α, then the odds are high.

Proof of (1). Start from the observation that

$$\begin{aligned}\mathrm{Cr}(\imath F \mid \imath E) &\geq \sum_r \mathrm{Cr}(\imath F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \imath E) \\ &= \sum_r \mathrm{Cr}(\imath F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \imath E \,\&\, \imath R)\end{aligned}$$

where the sum is over all non-negative rational r. In the second line, I introduced $\imath R$ using the assumption, stated as part of ADMISSIBILITY, that E entails R: by this assumption, $\imath E$ is equivalent to $\imath E \,\&\, \imath R$. I will consider each term in this sum, one at a time.

Recall the identity $\mathrm{Cr}(X \mid Y) = \mathrm{Cr}(X \,\&\, Y)/\mathrm{Cr}(Y)$. I will use this to repackage conditional probabilities without detailed explanation. (Instead of this identity, one could use the axioms for Popper functions.) As a first application, for each candidate value of r we have

$$\begin{aligned}\mathrm{Cr}(\imath F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \imath E \,\&\, \imath R) = \mathrm{Cr}(\imath E \mid \imath F \,\&\, \imath R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ \times \mathrm{Cr}(\imath F \mid \imath R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ \times \mathrm{Cr}(\mathrm{Rat}_R^{F/G} = r \mid \imath R)/\mathrm{Cr}(\imath E \mid \imath R).\end{aligned}$$

Next, applying the definition of λ to the first factor and applying INDIFFERENCE to the second, we find that the right-hand side is

$$\begin{aligned}\geq \lambda \times \mathrm{Cr}(\imath E \mid \imath G \,\&\, \imath R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ \times r \times \mathrm{Cr}(\imath G \mid \imath R \,\&\, \mathrm{Rat}_R^{F/G} = r) \\ \times \mathrm{Cr}(\mathrm{Rat}_R^{F/G} = r \mid \imath R)/\mathrm{Cr}(\imath E \mid \imath R).\end{aligned}$$

Repackaging the conditional probabilities, this is equal to

$$\lambda \times r \times \mathrm{Cr}(\mathrm{Rat}_R^{F/G} = r \mid \imath E \,\&\, \imath G) \times \mathrm{Cr}(\imath G \mid \imath E).$$

Call this expression $a(r)$. If we sum $a(r)$ over all values of r we get

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

So, all together,

$$\begin{aligned}\mathrm{Cr}(\iota F \mid \iota E) &\geq \sum_r \mathrm{Cr}(\iota F \,\&\, \mathrm{Rat}_R^{F/G} = r \mid \iota E) \\ &\geq \sum_r a(r) \\ &\geq \lambda \times \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G) \times \mathrm{Cr}(\iota G \mid \iota E).\end{aligned}$$

Dividing through by $\mathrm{Cr}(\iota G \mid \iota E)$ yields (1). □

Proof of (2). Note that R is automatically admissible with respect to R, F, G, and r. So, by the argument just given, we have INDIFFERENCE*, i.e.

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

(All we need to derive this is INDIFFERENCE.) Applying the definition of α to the left-hand side, and the definition of β to the right, we get

$$\frac{1}{\alpha} \times \mathrm{Odds}(\iota F / \iota G \mid \iota E) \geq \frac{1}{\beta} \times \mathbb{E}(\mathrm{Rat}_R^{F/G} \mid \iota E \,\&\, \iota G).$$

This rearranges to the inequality (2). □

References

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

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. (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.

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

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

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.

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.

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.

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