Another Patch for the Simulation Argument

Based on the rejected Oct 2016 Analysis submission titled

Errors in Bostrom/Kulczycki Simulation Arguments

Dr. R. Dustin Wehr

March 23, 2023

Abstract

Both of the previously "patched" versions of the Bostrom/Kulczycki ancestor simulation argument contain significant objective errors at the boundary between mathematics and gloss, specifically in the specification of the intended interpretation of extreme vague magnitude terms. The errors, with the parameter settings used by the authors to justify their striking glosses, allow formal deduction of absurdities such as

- Every "astronomically large factor" is smaller than 0.00000102.
- 1 is "vastly greater than" every number less than 989,901.

Whatever one's opinion on the value of Bostrom et al.'s simulation argument publications (e.g. with respect to epistemological assumptions), it is indisputable that the persuasiveness of the arguments benefits from these errors.

The errors were discovered while attempting to formalize the arguments as theorems in predicate logic, with the confounding factors of vagueness, subjectiveness, and uncertainty isolated as advocated in [Weh15]. This paper fixes the errors, as much as possible, proving optimal versions of the theorems. It provides a guide for readers to evaluate the theorems for themselves, and suggests sound settings of the theorem parameters that have *relatively* simple, accurate statements in English, free of vague magnitude terms, making them easier to properly understand and rigorously critique than the previous informal statements. Unfortunately, the corrected theorems do not have glosses with nearly the same punch as in Bostrom et al.'s publications; despite my

best efforts to simplify the parameters as generously as possible, the theorem statements remain rather technical and difficult to subjectively evaluate, since the parameters cannot be pushed to extremes that make any of the theorem's disjuncts clearly-compelling without making one of the other disjuncts obviously true and uninteresting (and thus trivializing the theorem).

Since Bostrom's simulation argument is among the most widely popularized arguments in Analytic Philosophy in the last several decades, I feel there should be much value in publishing a correct version, which doesn't prove more than it claims to prove.

Contents

Introduction

With the help of its wide online popular science distribution, Bostrom's Simulation Argument is a plausible candidate for the most popularized argument in the history of analytic philosophy. In 2016, I was looking for a well-known argument from analytic philosophy with which to promote the method of *interpreted formal proof dialogues* that I advocated for in my PhD thesis. The patched simulation arguments were the closest thing to mathematically rigorous philosophy that I was aware of at the time.[i] But the dialogue did not make it even to the second turn –where issues involving subjectiveness, vagueness, and uncertainty can begin to be addressed– because I found a mistake in the math during the first turn, while carrying out a first formalization of the argument in predicate logic. I notified Bostrom and Kulczycki (who disagreed about the significance) and submitted a manuscript with the results to *Analysis*, which had published their paper.

For the rest of the introduction, I graciously borrow an eloquent excerpt from one of the two anonymous reviews of that unsuccessful Analysis submission.[ii] It is better than I could do. Ellipses omit only references to sections and pages of the previous manuscript.

> *Overall, this is a very good paper that merits publication on the grounds that it corrects a flawed paper previously published in Analysis. I have some suggestions for minor revisions that should be taken into account.*
>
> *Overview of the paper*
>
> *Bostrom (2003) argued that at least one of the following three claims is true: (1) the fraction of civilizations that reach a 'post-human' stage is approximately zero; (2) the fraction of post-human civilizations interested in running 'significant numbers' of simulations of their own ancestors is approximately zero; (3)*

[i] I have since learned of the rigorous body of work on population axiology by Arrhenius and others; those are stronger examples of mathematical rigor.

[ii] Incidentally, a professor of analytic philosophy I spoke to informed me that two reviews is standard in philosophy, and a submission will normally be rejected if either review is even slightly negative. Then he baffled me, by saying that in the case of a critical paper, it is normal in philosophy to have one of those must-please anonymous reviews be written by an author of the criticized paper! Imagine that! Hopefully he was mistaken.

the fraction of observers with human-type experiences that are simulated is approximately one.

The informal argument for this three-part disjunction is that, given what we know about the physical limits of computation, a post-human civilization would be so technologically advanced that it could run 'hugely many' simulations of observers very easily, should it choose to do so, so that the falsity of (1) and (2) implies the truth of (3). However, this informal argument falls short of a formal proof.

Bostrom himself saw that his attempt at a formal proof in the (2003) paper was sloppy, and he attempted to put it right in Bostrom and Kulczycki (2011). The take-home message of . . . the manuscript under review is that these (2011) reformulations of the argument are still rather sloppy. For example, the author points out . . . that the main text of B&K inaccurately describes the mathematical argument in the appendix: the appendix uses an assumption much more favourable to B&K's desired conclusion than the assumption stated in the main text. Moreover, B&K's use of vague terms such as 'significant number' and 'astronomically large factor' creates a misleading impression. The author shows, amusingly, that the 'significant number' must be almost 1 million times greater than the 'astronomically large factor' for their argument to work. . .

. . . the author provides a new formulation of the simulation argument that is easily the most rigorous I have seen. This formulation deserves to be the reference point for future discussions of the argument's epistemological consequences.

Note 1. This paper is narrowly focused, and does not attempt to provide a summary of other kinds of criticisms of the simulation argument [Bir13] [Wea03] [Bru08] [Lew13], which might be more important than this quite technical one. Those critical works deserve summaries, but I see no incentive for me to put any more time into this than I already have; indeed, the significant time I

spent already has not benefited me at all. The rigour in contemporary analytic philosophy is very poor, and I do not think they are interested in having that pointed out by outsiders.

Also, throughout the paper, whenever possible I use via quotation the informal English glosses of [Bos03] and [BK10] for axioms, definitions and theorems, rather than introducing my own. In general, I take a completely uncritical stance on the nuanced epistemological issues involved, focusing, essentially, only on the math and logic.

This paper concerns the widely-publicized Simulation Argument, first published by Bostrom in [Bos03], and "patched" in two different ways years later in [BK10] to correct an error. Here, I correct three further serious errors, one in the Patch 1 version and two in the Patch 2 version. I also improve the formalization, give complete proofs, and demonstrate the significance of the errors. Once the result is corrected and formalized, I argue that any setting of the theorem's parameters, together with an impressive English gloss, would need to benefit from a severe instance of an equivocation pattern involving vague magnitude terms (e.g. "significant number of," "very likely," "extremely small," "almost certainly," "astronomically large"; see Sections 1.1, 2.1, and 3 in particular), which [Wal08] calls *variability of strictness of standards*. Variability of strictness of standards occurs when the intended interpretation of a (or several) vague predicate(s) with a natural 1-dimensional notion of magnitude (the "strictness") is left too vague, and the persuasiveness of the argument benefits from the informal intended interpretation using a strong/large-magnitude interpretation for some assumptions, and a weak/small-magnitude interpretation for others.[iii] In Section 3 I give new concise statements of the corrected simulation arguments which, for anyone concerned about rigour, should be used in preference to those in [BK10] and [Bos03].

This paper assumes familiarity with [BK10].

[iii] [Wal08] demonstrates the pattern by giving a proof of "Nobody should ever give married," which exploits the vagueness of two predicates: roughly, *person p can safely predict proposition A*, and *person p and person q are compatible*. For a formalization of the argument in predicate logic, and recommendation on how to methodically criticize it, see page 17 of [Weh15].

Throughout, ***PH*** *abbreviates "posthuman".*

Note 2. The following Definition 1 is given at a level above the formality of predicate logic. If given in a fully formal manner, it would define a set of first-order $\mathcal{L}$-structures for a particular language $\mathcal{L}$, which includes not only the five symbols $\{\mathsf{C}, \mathsf{C}^{\mathsf{PH}}, \mathsf{pop}, \#\mathsf{sims}, N, \mathsf{count}^E\}$ that are particular to the Simulation Argument (and a couple more for the Patch 2 version), but also many mathematical symbols that have standard meanings, which would be fixed in the definition of the set of $\mathcal{L}$-structures. I will save the reader from excessive jargon and pedantry by counting on our shared understanding of the standard meanings of symbols for numbers and finite sets.

Definition 1. A Simulation Argument 1-model[iv], or just 1-model, is given by:

- A finite nonempty set C, for the "human-level technological civilizations" [BK10]. I use the clearer term *advanced human-like civilizations*, where *advanced* means as advanced as the current state of the human race.
- A possibly-empty subset C^{PH} of C, for the civilizations that eventually reach a posthuman (PH) stage.
- For each $c \in \mathsf{C}$, a natural number $\mathsf{pop}(c)$, for the cumulative pre-PH population size[v] of c.
- For each $c \in \mathsf{C}^{\mathsf{PH}}$, a natural number $\#\mathsf{sims}(c)$, for the number of pre-PH-phase ancestral simulations that c does in its PH phase.
- A positive integer N, for the number of ancestor simulations that a civilization must eventually run in order to be considered to have run "a significant number of ancestor simulations" [BK10].[vi]

Whenever a model is fixed, we also have the following abbreviations:

- $\mathsf{C}^{\overline{\mathsf{PH}}} \coloneqq \mathsf{C}/\mathsf{C}^{\mathsf{PH}}$, the civilizations that never reach a PH stage.

[iv] In contrast to 2-model, defined in the Section 2.

[v] The cumulative pre-posthuman population size is the number of (real) beings that lived before the time when the civilization was able to run phenomenally convincing ancestor simulations.

[vi] Intuitively it is an argument parameter, like the parameters q_1, q_2, q_3, d introduced later, but for the purpose of Theorem 1 I make it part of the model.

- $\mathsf{C}_{<N} \coloneqq \{c \in \mathsf{C} \mid \#\mathsf{sims}(c) < N\}$ and $\mathsf{C}_{\geqslant N} \coloneqq \{c \in \mathsf{C} \mid \#\mathsf{sims}(c) \geqslant N\}$, the civilizations that run fewer than N and at least N ancestor simulations, respectively.
- $\#U = \sum_{c \in \mathsf{C}} \mathsf{pop}(c)$ is the total number of unsimulated observers.[vii]
- $\#S \coloneqq \sum_{c \in \mathsf{C}^{\mathsf{PH}}} \mathsf{pop}(c) \cdot \#\mathsf{sims}(c)$ is the total number of simulated observers. The formula conveys [BK10]'s intended interpretation of "ancestor simulation," in which each of a civilization c's simulations contains exactly as many simulated observers as there were unsimulated observers in the pre-PH phase of c.[viii] This formula is one of the reasons that the argument specifies *ancestor* simulations instead of some more general category of simulations of intelligent beings. It is one of the main novelties of Bostrom's argument.
- $\mathsf{avgpop}_{<N} \coloneqq \left(\sum_{c \in \mathsf{C}_{<N}} \mathsf{pop}(c) \right) / |\mathsf{C}_{<N}|$ and $\mathsf{avgpop}_{\geqslant N} \coloneqq \left(\sum_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c) \right) / |\mathsf{C}_{\geqslant N}|$ are the average number of unsimulated observers in the pre-PH phase of civilizations in $\mathsf{C}_{<N}$ and $\mathsf{C}_{\geqslant N}$, respectively.

Next, we introduce some symbols and definitions from [BK10], though with different notation that meshes better with the other notation used in this paper.

Note that all quotations in the following three definitions are from that paper.

Definition 2 ($f_{\mathrm{PH}}, f_{\overline{\mathrm{PH}}}, q_1, \mathsf{Prop\ 1}$)**.** f_{PH} is informally defined as "The fraction of human-level civilizations that reached a posthuman stage." It and $f_{\overline{\mathrm{PH}}}$ are uncontroversially defined formally by:

$$f_{\mathrm{PH}} \coloneqq \frac{|\mathsf{C}^{\mathsf{PH}}|}{|\mathsf{C}^{\mathsf{PH}}| + |\mathsf{C}^{\overline{\mathsf{PH}}}|} \qquad f_{\overline{\mathrm{PH}}} \coloneqq 1 - f_{\mathrm{PH}}$$

$\mathsf{Prop\ 1}$ is intended to express "The human species is very likely to go extinct before reaching a PH stage."

$$f_{\mathrm{PH}} < q_1$$

where q_1 is a $[0, 1]$ parameter of the argument. In the appendix of [BK10], an example proof with $q_1 = .01$ is demonstrated.

[vii] Note the implicit assumption that all civilizations are non-overlapping.

[viii] Note the implicit assumption that all ancestor simulations are non-overlapping.

Definition 3 ($f_{\geqslant N}, q_2, \mathsf{Prop\ 2}$)**.** $f_{\geqslant N}$ is informally defined as "The fraction of posthuman civilizations that are interested in running a significant number of ancestor simulations". Here, "significant number" means N. It is uncontroversially defined by

$$f_{\geqslant N} \coloneqq \frac{|\mathsf{C}_{\geqslant N}|}{|\mathsf{C}^{\mathsf{PH}}|}$$

Prop 2 is intended to express that $f_{\geqslant N}$ is "extremely small." Formally:

$$f_{\geqslant N} < q_2$$

where q_2 is another $[0, 1]$ argument parameter, set to .01 in [BK10]'s example proof.

Definition 4 ($f_{\text{sim}}, q_3, \mathsf{Prop\ 3}$)**.** f_{sim} is informally defined as "...the fraction of all observers in the universe with human-type experiences that are living in computer simulations." It is uncontroversially defined by

$$f_{\text{sim}} \coloneqq \frac{\#S}{\#S + \#U}$$

Prop 3 is intended to express "We are almost certainly living in a computer simulation." Formally:

$$f_{\text{sim}} > q_3$$

where q_3 is another $[0, 1]$ parameter, set to .99 in [BK10]'s example proof.

1 First Patch and Error 1

Patch 1 is described in [BK10] as:

> ...a very weak assumption to the effect that the typical duration (or more precisely, the typical cumulative population) of the pre-posthuman phase *does not differ by an astronomically large factor* between civilizations that never run a significant number of ancestor simulations and those that eventually do. For example, in an appendix we show how by assuming that the difference is no greater than a factor of one million we can derive the key tripartite disjunction.

Formally:

Definition 5 (Patch 1)**.**

$$\frac{\mathsf{avgpop}_{<N}}{\mathsf{avgpop}_{\geqslant N}} \leqslant d \text{ [ix]}$$

where d is a natural number parameter of the argument. The previous quoted passage tells us that the appendix uses the parameter setting[x]:

$$\frac{\mathsf{avgpop}_{<N}}{\mathsf{avgpop}_{\geqslant N}} \leqslant 1 \text{ million} \tag{1}$$

Unfortunately, rather than inequality (1) above, the appendix of [BK10] erroneously uses the much stronger[xi] assumption:

$$\frac{\mathsf{avgpop}_{<N}}{\mathsf{avgpop}_{\geqslant N}} \leqslant \frac{N}{1 \text{ million}} \tag{2}$$

Then, under the additional, quite reasonable assumption $N \geqslant 9900$,[xii] they proved that

$$\mathsf{Patch\ 1} \rightarrow \mathsf{Prop\ 1} \vee \mathsf{Prop\ 2} \vee \mathsf{Prop\ 3}$$

However, a massively larger lower bound on N, close to one trillion (see Corollary on page 11), is needed when the erroneous (2) is replaced with (1). This is Error 1.

If you are not interested in the details, skip to Section 1.1 on page 11 now.

It turns out to be easier to state and understand the dependency of the argument on its parameters N, d, q_1, q_2, q_3 if we do a change of variables. We restrict our attention to settings of the parameters where q_1, q_2, q_3 are all in $(0, 1)$,[xiii] and replace them with $\mathbb{R}^+$

[ix] Some models have $\mathsf{avgpop}_{\geqslant N} = 0$, in which case the left side is undefined and so the inequality is false, and then all the results in this paper that depend on Patch 1 hold trivially.

[x] Technically there is exactly one other a priori reasonable interpretation of "the difference is no greater than a factor of one million," in which the numerator and denominator of the left hand side of the inequality are flipped. However, that alternative can be ruled out from other statements in the paper, and in any case could not have been intended since it does not help to prove the trilemma.

[xi] Technically, (2) is only stronger than (1) when $N \leqslant 10^{12}$. But such a large value of N trivializes the theorem and does not fit the gloss "significant number" used for N, so we can assume that was not intended.

[xii] Reasonable with respect to the given intended interpretation that a "significant number of" ancestor simulations means $\geqslant N$ ancestor simulations.

[xiii] i.e. none are 1 or 0. This does not affect the criticism.

parameters:

$$k_1 = \frac{1-q_1}{q_1} \qquad k_2 = \frac{1-q_2}{q_2} \qquad k_3 = \frac{q_3}{1-q_3}$$

and then the reader can check:

$$\begin{aligned} \mathsf{Prop\ 1} &\equiv |\mathsf{C}^{\overline{\mathsf{PH}}}| > k_1|\mathsf{C}^{\mathsf{PH}}| \\ \mathsf{Prop\ 2} &\equiv |\mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N}| > k_2|\mathsf{C}_{\geqslant N}| \\ \mathsf{Prop\ 3} &\equiv \#S > k_3 \#U \end{aligned}$$

For example, the parameter setting $q_1 = .01, q_2 = .01, q_3 = .99$ used in [BK10] corresponds to $k_1 = k_2 = k_3 = 99$. Let $\vec{k}$ abbreviate k_1, k_2, k_3.

Define the expression (LB for lower bound):

$$\mathsf{LB}(d, \vec{k}) \coloneqq dk_3(k_1 + k_2 + k_1k_2) + k_3$$

Now we can state the main results for Patch 1. Let $\models_1$ denote entailment for 1-models (Definition 1).[xiv]

Theorem 1. *For all settings of the parameters $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$:*

$$N > \mathit{LB}(d, \vec{k}), \mathit{Patch\ 1} \models_1 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3}$$

The following companion theorem shows that the lowerbound on N in Theorem 1 cannot be weakened, and so in that sense Theorem 1 is as strong as possible given the other assumptions.

Theorem 2. [xv] *For all settings of the parameters $d \in \mathbb{N}$, $\vec{k} \in (\mathbb{N}^+)^3$:*

$$N \geqslant \mathit{LB}(d, \vec{k}), \mathit{Patch\ 1} \not\models_1 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3}$$

For example, if we fix the parameters q_1, q_2, q_3, d as they are (or should be, in the case of d) in the appendix of [BK10], then we get:[xvi]

[xiv] That is, if A is a sentence and Γ a set of sentences, then $\Gamma \models_1 A$ means every 1-model that satisfies each sentence in Γ must also satisfy A.

[xv] It is not a mistake that the domain of $\vec{k}$ is different here. The proof of Theorem 2 (5.3.1), as it is now, uses the fact that k is an integer, whereas the proof of Theorem 1 does not. Thanks to an anonymous referee for pointing out that this deserves explanation, since it looks like it could be an error.

[xvi] Actually what one gets from substitution is $989,901,000,099$; we round in the sound direction for both statements of the Corollary.

Corollary. *If $d = 1$ million, $q_1 = .01, q_2 = .01, q_3 = .99$,*[xvii] *then*

$$N > 0.99 \text{ trillion}, \textsf{Patch 1} \models_1 \textsf{Prop 1} \vee \textsf{Prop 2} \vee \textsf{Prop 3}$$

$$N \geqslant 0.989 \text{ trillion}, \textsf{Patch 1} \not\models_1 \textsf{Prop 1} \vee \textsf{Prop 2} \vee \textsf{Prop 3}$$

1.1 Effect on the argument

Recall the prose definitions of Prop 2 and Patch 1 from [BK10]:

> Prop 2: The fraction of posthuman civilizations that are interested in running a significant number of ancestor simulations is extremely small.
>
> Patch 1: ...the typical cumulative population... of the pre-posthuman phase *does not differ by an astronomically large factor* between civilizations that never run a significant number of ancestor simulations and those that eventually do.

Recall that "astronomically large factor" means d, and "significant number" means N, which we now know must be larger than $dk_3(k_1 + k_2 + k_1k_2) + k_3$ where the magnitudes of k_1, k_2, and k_3 should be chosen to reflect the severity of the terms "very likely," "extremely small," and "almost certainly," respectively.

To see the significance of the error, I consider the effect on the one fully-fleshed out proof given in [BK10]. There the partial parameter setting $k_1 = k_2 = k_3 = 99$ (equivalently $q_1 = .01, q_2 = .01, q_3 = .99$) is used. Although I believe that the use of vague magnitude terms should in general be avoided when giving interpretations of proofs, their use for these parameters is harmless enough:

> $1 - q_1$ = probability 0.99 interpreted as "very likely"
>
> q_2 = probability 0.01 interpreted as "extremely small"
>
> q_3 = probability 0.99 interpreted as "almost certainly"

But when we consider the meaning of the lower bound on N now, we see a serious problem.

[xvii] i.e. $k_1 = k_2 = k_3 = 99$

Substituting [BK10]'s settings of $\vec{k}$ into $N > dk_3(k_1 + k_2 + k_1k_2) + k_3$, we get:

$$N > d \times 989,901$$

And then, substituting in the the intended interpretations of N and d, we get:

$$\text{“significant number”} > \text{“astronomically large factor”} \times 989,901$$

Thus, the persuasiveness of [BK10] benefits from what is clearly a wildly misleading definition of “significant number.”

2 Second Patch and Error 2

The Patch 2 argument introduces the idea of *E-observers*, which are the observers (unsimulated and simulated) that satisfy a chosen fixed predicate E. You, by appropriate choice of the predicate E, are an E-observer. The *bland indifference principle*[Bos03] of the Patch 1 argument, from which the authors of [BK10] justify the jump

$$\begin{aligned} &f_{\text{sim}} \text{ fraction of observers are simulated} \\ \rightarrow\ &\text{you should believe with credence } f_{\text{sim}} \text{ that you are simulated} \end{aligned}$$

is made dependent on E: since you cannot tell whether or not you are a simulated E-observer, you should believe, with credence equal to the fraction of simulated E-observers over all E-observers, that you are simulated. As with the Patch 1 Simulation Agument from the previous section, I will accept the qualitative assumptions of the Patch 2 argument uncritically, focusing only on the mathematical aspects.

The Patch 2 argument was not fleshed out in [BK10]. I do that here, and it turns out that the mathematics of the Patch 1 and Patch 2 arguments are practically the same (and the proof of Corollary 1 on page 30 follows easily from Theorem 1).

The (unfixed and fundamental[xviii]) language of the Patch 2 argument is the language of the Patch 1 argument $\{\mathsf{C}, \mathsf{C}^{\mathsf{PH}}, \mathsf{pop}, \#\mathsf{sims}, N\}$ plus the new symbol count^E.

[xviii] e.g. the symbol / for set difference is fixed, and the symbol $\mathsf{C}_{<N}$ is defined in terms of fixed and fundamental symbols, and thus is not itself fundamental. See Note 2 (pg 6).

Definition 6. A Simulation Argument 2-model, or just 2-model, is a 1-model together with a function count^E that counts the number of E-observers in any given civilization.

When a 2-model is fixed, we also use the following abbreviations:

- $\mathsf{avgpop}^E_{\geqslant N}$ and $\mathsf{avgpop}^E_{<N}$, the average number of E-observers in civilizations from $\mathsf{C}_{\geqslant N}$ and $\mathsf{C}_{<N}$, respectively. That is, $\mathsf{avgpop}^E_{\geqslant N} = \left[\sum_{c\in\mathsf{C}_{\geqslant N}} \mathsf{count}^E(c)\right] / |\mathsf{C}_{\geqslant N}|$.
- $\mathsf{C}^{E\geqslant 1}_{<N}$ and $\mathsf{C}^{E\geqslant 1}_{\geqslant N}$, the civilizations in $\mathsf{C}_{<N}$ (resp. $\mathsf{C}_{\geqslant N}$) that contain at least one E-observer.
- $\#U^E \coloneqq \sum_{c\in\mathsf{C}} \mathsf{count}^E(c)$, the total number of **u**nsimulated E-observers.
- $\#S^E \coloneqq \sum_{c\in\mathsf{C}^{\mathsf{PH}}} \mathsf{count}^E(c)\cdot\#\mathsf{sims}(c)$, the total number of **s**imulated E-observers.

Definition 7 ($f^E_{\mathrm{sim}}, q_3, k_3, \mathsf{Prop}\ 3'$)**.** The role of f_{sim} in the previous argument is played by

$$f^E_{\mathrm{sim}} \coloneqq \frac{\#S^E}{\#S^E + \#U^E}$$

The role of $\mathsf{Prop}\ 3$ in the previous argument is played by $\mathsf{Prop}\ 3'$, defined by

$$f^E_{\mathrm{sim}} > q_3$$

Or equivalently $\#S^E > k_3\#U^E$ using the alternate parameterization introduced in the previous section, which we use in this section as well.

The informal definition of $\mathsf{Patch}\ 2$ is provided by the following quote from [BK10] (page 4):

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

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

Unfortunately, that does not quite suffice, which brings us to Error 2 of [BK10], which when fixed results in the absurdity explained in Section 2.1. We actually need the fraction mentioned in (i), which is $|\mathsf{C}^{E\geqslant 1}_{\geqslant N}|/|\mathsf{C}_{\geqslant N}|$, to be "substantial" *relative to* the corresponding fraction for civilizations that run fewer than N ancestor simulations, where the meaning of "substantial" is an unnamed argument parameter. There is also the unnamed parameter that defines (ii)'s "vastly greater than." Fortunately, it is only the product of those parameters that matters for stating the following theorems, so our version of Patch 2, Definition 8, collapses them into one parameter. **I will use d for this parameter, due to the very similar role it plays to the d used in Section 1 for the Patch 1 argument.** Temporarily adopting the notation from the previous quote, we are using these facts:

$$\left(\frac{|\mathsf{C}^{E\geqslant 1}_{\geqslant N}|}{|\mathsf{C}_{\geqslant N}|}\right)\cdot H_s(E) = \mathsf{avgpop}^E_{\geqslant N} \qquad \text{and} \qquad \left(\frac{|\mathsf{C}^{E\geqslant 1}_{<N}|}{|\mathsf{C}_{<N}|}\right)\cdot H_n(E) = \mathsf{avgpop}^E_{<N}$$

Definition 8 (Patch 2)**.**

$$\frac{\mathsf{avgpop}^E_{<N}}{\mathsf{avgpop}^E_{\geqslant N}} \leqslant d$$

An important special case of the Patch 2 argument, used in [BK10], is clarified by the following:

Fact 1. *If E is such that no civilization has more than one E-observer, then Patch 2 is equivalent to*

$$\frac{|C^{E\geqslant 1}_{<N}|/|C_{<N}|}{|C^{E\geqslant 1}_{\geqslant N}|/|C_{\geqslant N}|} \leqslant d$$

That is, the fraction of $\mathsf{C}_{<N}$ civilizations with an E-observer is at most d times larger than the fraction of $\mathsf{C}_{\geqslant N}$ civilizations with an E-observer.

You may have wondered if whether we are counting all observers, or only E-observers, is inconsequential, since the E-restriction is applied to all civilizations. One way of formalizing

that intuition involves proving that the 2-models of the Patch 2 argument can be mapped (in a suitable truth-preserving way) to the 1-models of the Patch 1 argument, where the E-observers of the former become the observers of the latter.[xix] That is the approach we take in the proof of Corollary 1 (page 30).

Let $\models_2$ denote entailment with respect to Definition 6.

Corollary 1. *For all settings of the parameters* $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$*:*

$$N > \mathit{LB}(d, \vec{k}), \mathit{Patch\ 2} \models_2 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3'}$$

Corollary 2. *For all settings of the parameters* $d \in \mathbb{N}, \vec{k} \in (\mathbb{N}^+)^3$*:*

$$N \geqslant \mathit{LB}(d, \vec{k}), \mathit{Patch\ 2} \not\models_2 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3'}$$

Note 3. The Patch 1 argument is a special case of the Patch 2 argument; just take E to be *true*.

I have gone to some trouble to formalize the Patch 2 argument results in such a way that Corollary 1 and Corollary 2 are almost identical to Theorem 1 and Theorem 2, as it makes the proofs of Corollary 1 (page 30) and Corollary 2 (page 33) easier.

2.1 Effect on the argument

The English gloss of [BK10]'s Patch 2 result suffers from a problem similar to that of the English gloss of their Patch 1 result, though it takes more effort to show this since the authors provide only a sketch of the Patch 2 argument. Recall that in Section 1.1 we showed that

[xix] **Pedantic note:** In the case of [BK10]'s example where E is "My computer age birth rank is 1 billion," where every civilization in a 2-model has 0 or 1 E-observers (as in Fact 1), the corresponding 1-model civilizations have cumulative population size 0 or 1. Of course, a civilization with no observers is probably not compatible with what the authors of [BK10] had in mind by "human-level technological civilizations" [BK10]. *However*, models containing such trivial civilizations are acceptable according to the assumptions *needed* in order to prove the mathematical statement Theorem 1. Even if the definitions were tightened to exclude civilizations with no observers, we would still be able to use Theorem 1 to prove Corollary 1. In fact, in the proofs in the appendix, we make use of the permissibility of mathematical models of civilizations with zero cumulative pre-PH population size that nonetheless do pre-PH ancestor simulations. I also give further explanation of why such reasoning is beyond reproach.

their assumptions, in relation to their assigned informal English interpretations, imply

$$\text{“significant number”} > \text{“astronomically large factor”} \times 989,901$$

Recall from our restatement on page 13 of [BK10]'s gloss of their version of Patch 2: “It is *not* the case that $H_n(E)$ is vastly greater than $H_s(E)$.”

It turns out that, by their word usage and suggested parameter settings:

“vastly greater than” means greater than by a factor of

$$\frac{\text{“significant number”}}{989{,}901}$$

I leave deriving the previous absurdity as an exercise for the reader, with this tip: Examine Definition 8 and the two equations that precede it – importantly, the objective error in [BK10]'s Patch 2 argument sketch must be fixed *before* deriving the absurdity.

A second serious error in the Patch 2 argument is presented in Section 3.

3 Improved formalization and Guide to Evaluating the Simulation Argument for Yourself

The reader may wonder why I have gone through so much trouble to give optimized versions of the Simulation Arguments in terms of the function LB. It is for two reasons. First, I wanted to be sure that I was treating not just the published form of Bostrom/Kulczycki's arguments fairly, but also the ideas behind that form. Thus, I did not settle for merely demonstrating absurd consequences of the errors in the arguments, as I did in Sections 1.1 and 2.1, since such a thing could in principle be fixed; I also proved theorems that attempt to characterize the limitations of the ideas used in the arguments (Theorem 2 and Corollary 2). The second reason is that we will use, in this section, settings of the parameters that require knowing the exact form of LB.

Observation: Whether or not you have extra concerns about the assumptions and interpretation of the Patch 2 argument that you don't have about the Patch 1 argument, without loss of generality we may restrict attention to the Patch 2 argument. The reason

is, first of all, as explained in Note 3, the corrected Patch 1 argument is mathematically a special case of the corrected Patch 2 argument. Second, the next subsection will delay asking the reader to limit their choices for E until Step 3, so a reader who prefers the Patch 1 argument can stop just before then, and use $E = \mathit{true}$ to reduce to the Patch 1 argument, instead of the E I recommend (which is a strengthening of [BK10]'s example E).

With that observation in mind, let us first take what we have learned about the corrected Simulation Argument to give an equivalent, concise statement that is free from problematic vague magnitude terms such as "significant number of" and "astronomically large." After that, I will try to persuade the reader to fix a couple of the parameters, to get a simpler form.

Theorem 3. *Let d, N be natural numbers and $\vec{k}$ a triple of positive real numbers. Define four propositions:*

Prop 1: There are more than k_1 times more advanced-human-like civilizations that never reach the PH stage than there are that eventually reach the PH stage.

Prop 2: Among the advanced human-like civilizations that eventually reach the PH stage, the number that run fewer than N ancestor simulations is more than k_2 times greater than the number that run at least N ancestor simulations.

Prop 3: There are more than k_3 times more simulated E-observers than non-simulated E-observers.

Prop 4[xx]*: The average number of E-observers in advanced human-like civilizations that run fewer than N ancestor simulations is more than d times larger than the average number of E-observers in human-like civilizations that run at least N simulations.*

Let $\models_2$ denote entailment with respect to Definition 6. Then

$$N > dk_3(k_1 + k_2 + k_1k_2) + k_3 \models_2 \textit{Prop 1} \vee \textit{Prop 2} \vee \textit{Prop 3} \vee \textit{Prop 4}$$

[xx] Formerly ¬Patch 2.

and when the $\vec{k}$ are integers[xxi]*, the bound is tight:*

$$N \geqslant dk_3(k_1 + k_2 + k_1k_2) + k_3 \not\models_2 \textit{Prop 1} \vee \textit{Prop 2} \vee \textit{Prop 3} \vee \textit{Prop 4}$$

We will now be able to see more clearly the subtle difficulty of evaluating the simulation argument:

When $k_1k_2k_3$[xxii] is large, as suggested in [BK10], the truth (or probable truth) of the less-interesting Prop 2 and Prop 4 are difficult to assess, and moreover:

- Making d large makes Prop 4 less likely and Prop 2 more likely.
- Making d small makes Prop 2 less likely and Prop 4 more likely.

But we need both Prop 4 and Prop 2 to be unlikely in order to conclude that the more exciting proposition, Prop 1 $\vee$ Prop 3, is likely. This is the sense in which the impressiveness of the Simulation Arguments benefits from variability of strictness of standards, as mentioned in the introduction.

3.1 A Special Case of the Main Theorem to Focus On

Let A-Civilizations abbreviate "advanced, our-technology-or-better human-like civilizations".

Recall that none of the allowed settings of the argument parameters are *wrong*. Nor are they subjective; they simply yield different theorems, some of which you will find more interesting than others.

In this section, **I suggest ways of constraining the parameters that lead to simpler, and thus easier to criticize, English glosses, subject to the assumption that we are most interested in seeing when the two more dramatic propositions, Prop 1(doom) and Prop 3(you are probably simulated), are likely to be true.** Due to the interdependence of the propositions, making Prop 2 and Prop 4 less likely makes Prop 1 and Prop 3 more likely, but less dramatic, and so there is a trade off. Thus, our tactic is this: Make Prop 1 and Prop 3 as likely as possible *subject to the constraint* that they remain both *profound* and *easy to fully understand / grasp the implications of*.

[xxi] Probably unnecessary, but it's a current limitation of the proof.

[xxii] Note that $dk_1k_2k_3$ is the dominating term in $\mathsf{LB}(d, \vec{k})$

Step 1

Set $k_3 = 1$, so that Prop 3 becomes "There are more simulated E-observers than non-simulated E-observers," **eliminating one of the vague magnitude terms, while simultaneously lowering the needed lower bound on N, which makes Prop 2 more likely false** (which we want). This is optimal for a skeptic of the simulation argument, since weakening one of the possible conclusions from "You are very likely simulated" to "You are most likely simulated" hardly affects the boldness of Prop 1 $\wedge$ Prop 3, while making it harder to claim that Prop 2 is true.

If you accept [BK10]'s interpretation of the proposition using the "bland indifference principle", **then** a perfectly accurate gloss of Prop 3 is:

> Prop 3: You are more likely simulated than unsimulated.

In contrast, [BK10] effectively used $k_3 = 99$ in their illustrative example, but that likely only because they mistakenly thought they had a lot of slack to work with.

Step 2

Set k_1 *just* large enough to strongly overestimate your best-guess subjective probability that the human race is destroyed before reaching the PH stage[xxiii]. [BK10]'s suggestion of $k_1 = 99$ should suffice for all but the most pessimistic among us, but a smaller value may suffice as well. Keep in mind that a larger value of k_1 (or k_2, k_3, or d) weakens the Simulation Argument by forcing a larger value of N. For $k_1 = 99$, *if you accept [BK10]'s reasoning that we should, roughly, treat our civilization as a random sample from the A-Civilizations*, **then** a perfectly accurate gloss of Prop 1 is:

> Prop 1: There's at least a 99% chance that the human race is destroyed before it reaches the posthuman stage.

[xxiii] Recall that "the PH stage" does not merely mean super-intelligence. It demands all the advances necessary to allow for an ancestor simulation that is as convincing as the world we live in now.

Step 3

Take E to be an elaboration of a proposition that, first of all, like [BK10]'s "My computer age birth rank is 1 billion", singles you out within the human race, and singles out at most 1 entity in each of the A-Civilizations (so that the average number of E-observers in any set of A-Civilizations is at most 1). Second, your E should contain any knowledge about our world that is pertinent to whether we might be living in a simulation, **or pertinent to whether we are living in a civilization that dies out before the PH stage or does fewer than N ancestor simulations.** The bolded point is because that question is not independent, in the Bayesian probability sense, of the question of whether we are living in a simulation. For example, if we are living in a civilization that dies out before the PH state, then we are definitely not living in an accurate ancestor simulation. Thus, "My computer age birth rank is X" is not sufficient." **You *must* at least strengthen E to something like the following**:

> My computer age birth rank is X, I live in a single-planet civilization, with several nations that have thermonuclear arsenals capable of destroying the civilization, (something about the state of affairs with climate change), (likewise for artificial intelligence existential risk), (likewise for pandemics), etc.

Note that the authors of [BK10] were aware of this issue, as evidenced by following quote:

> "As we attempt to convey with our suggested (start at a definition of) E above, any grounds for thinking that we are in a civilization with a significant chance of destroying itself before reaching the posthuman state, is grounds for thinking that E-observers are less likely to be in simulations."

Moving on, regardless of the specific definition of E, restricting E to definitions that single out at most 1 entity in every A-Civilization[xxiv] simplifies Prop 4 to:

> Prop 4 version (a): Let $\mathsf{C}_{<N}$ and $\mathsf{C}_{\geqslant N}$ be the A-Civilizations that run fewer than N and at least N simulations, respectively. The fraction of E-observer-having

[xxiv] Our-technology-or-better human-like civilizations

civilizations among $\mathsf{C}_{<N}$ (*including those that never make it to the posthuman stage*) is more than d times greater than the fraction of E-observer-having civilizations among $\mathsf{C}_{\geqslant N}$.

Optional Step 4

[BK10] argues that we can safely conclude that the previous simplified version of Prop 4 with is false, for any $d \geqslant 1$, *for their example* E, which would take us back to a tripartite disjunction. Their argument, which is *somewhat* plausible for their example E "my computer-age birth rank is X", is that "the frequency of E-observer-having civilizations among A-Civilizations that run at least N ancestor simulations" is 1. However, their example E ignores a ton of relevant evidence, since some A-Civilizations that make it to the posthuman stage never find themselves in as precarious a position as we appear to be now. For the E I sketched above (Step 3 on page 20), we cannot safely conclude Prop 4 is false unless we make d rather large. I will use $d = 100$, but of course feel free to choose your own.

Then Prop 4 further simplifies to:

> Prop 4 version (b): Let $\mathsf{C}_{<N}$ and $\mathsf{C}_{\geqslant N}$ be the A-Civilizations that run fewer than N and at least N simulations, respectively. The frequency of E-observer-having civilizations among $\mathsf{C}_{<N}$ (*including those that never make it to the posthuman stage*) is more than 100 times greater than the frequency of E-observer-having civilizations among $\mathsf{C}_{\geqslant N}$.

Moving on, we now have an argument with a single parameter, k_2. The final result of these simplifications is given on the next page.

To skeptics of the simulation argument, this is the most compelling single parameter **correct** *version of the fixed simulation argument that has been written.* If you wish to consider two parameter versions, which I believe benefits the Simulation Argument, I recommend stopping the simplifications before Optional Step 4.

A 0-parameter version, setting $k_2 = 100$, is given in Section 3.1.2.

3.1.1 Example Implication form, 1-Parameter Simplified English Version

I find the result is easier to understand as an implication. Instead of Prop 1 $\vee$ Prop 2 $\vee$ Prop 3 $\vee$ Prop 4, I put the negations of the more-technical Prop 2 and Prop 4 into the hypothesis, and the clearly-profound disjunction Prop 1 $\vee$ Prop 3 into the conclusion.

Let k_2 be any positive real number. See Step 3 for the meaning of "E-observer".

If

> ¬Prop 2: Among the advanced human-like civilizations that eventually reach the posthuman stage, there are less than k_2 times as many that run fewer than $N = 10000k_2 + 9901$ ancestor simulations than there are that run at least N ancestor simulations,

and[xxv]

> ¬Prop 4: Let $\mathsf{C}_{<N}$ and $\mathsf{C}_{\geqslant N}$ be the advanced human-like civilizations that run fewer than N and at least N simulations, respectively. The frequency of E-observer-having civilizations among $\mathsf{C}_{<N}$ (*including those that never make it to the posthuman stage*) is at most 100 times greater than the frequency of E-observer-having civilizations among $\mathsf{C}_{\geqslant N}$.

then

> At least one of Prop 1 or Prop 3 are true ("There's at least a 99% chance that the human race is destroyed before it reaches the posthuman stage", or "You are probably simulated").

Observe that the dependence of N on k_2 makes for a subtle task of setting k_2 to maximize the likelihood of ¬Prop 2 and ¬Prop 4 in the antecedent. Making k_2 large (say, 99 as in [BK10]) seems at first like a good strategy of falsifying Prop 2 in the antecedent, but then you notice that doing so raises N, and so **shifts more of the**

[xxv] Cross the following second hypothesis out if you are satisfied that Prop 4 is false **for your choice of** E.

eventually-posthuman civilizations into the category of running fewer than N ancestor simulations, which could shift credence in Prop 2 in either direction depending on the distribution of $\#\mathsf{sims}(\cdot)$, and can shift credence in Prop 4 in either direction depending on E. If Prop 2 is true then the implication is trivial and we conclude nothing about the profound disjunction Prop 1 $\vee$ Prop 3.

I suggest that the reader consider the theorem with [BK10]'s preferred $k_2 = 99$ before moving on to the next page where I use $k_2 = 10$.

3.1.2 Example Implication form, 0-Parameter Simplified English Version

Here we do a final simplification, fixing the parameter k_2 to 10. I also move the "doom" proposition from the consequent to the antecedent, so that the results tells us under what conditions we should expect that we are simulated.

To get the Patch 1 argument, replace "frequency of fraught" with "average cumulative population of".

Recall that one can take the definition of E-observer to be the trivial "any member of the civilization", which reduces the Patch 2 argument to the Patch 1 argument. I repeat my sketch of a suggested definition of "E-observer" from Step 3 here:

> My computer age birth rank is (whatever yours is), I live in a single-planet civilization, with several nations that have thermonuclear arsenals capable of destroying the civilization, (something about the state of affairs with climate change), (likewise for artificial intelligence existential risk), (likewise for pandemics), etc.

I abbreviate that further in the following. Rather than "advanced human-like civilization that have an E-observer", I say "fraught civilizations".

Suppose that

- **among the civilizations**[xxvi] **that eventually reach the posthuman stage, the number that run at least** 109,901 **(hereafter "a lot") ancestor simulations is at least** 10% **of the number that run less than that many,**[xxvii] *and*
- **the frequency of fraught civilizations among the advanced human-like civilizations that don't run a lot of ancestor simulations,** *including those that never make it to the posthuman stage*, **is at most 100 times greater than the frequency of fraught civilizations among the eventually-posthuman civilizations that run a lot of ancestor simulations,**[xxviii] *and*

[xxvi] advanced human-like civilizations
[xxvii] This is the instantiated ¬Prop 2
[xxviii] ¬Prop 4

- **there's less than a 99% chance that the human race is destroyed before it reaches the posthuman stage**[xxix]

then you are probably simulated.[xxx]

4 Conclusion

I corrected remaining errors in both "patched" versions of the Simulation Argument, and analyzed their significance carefully in the language of mathematical logic. I found that, although the corrected arguments are sound, their meaning is subtly dependent on the settings of (interdependent) parameters, and I do not believe the parameters can be set in a way that makes the arguments nearly as impressive as they appeared in [BK10].

References

[Bir13] Jonathan Birch. On the "simulation argument" and selective scepticism. *Erkenntnis*, 78(1):95–107, 2013.

[BK10] Nick Bostrom and Marcin Kulczycki. A patch for the simulation argument. *Analysis*, pages 54–61, 2010.

[Bos03] Nick Bostrom. Are we living in a computer simulation? *The Philosophical Quarterly*, 53(211):243–255, 2003.

[Bru08] Anthony Brueckner. The simulation argument again. *Analysis*, 68(3):224–226, 2008.

[Lew13] Peter J Lewis. The doomsday argument and the simulation argument. *Synthese*, 190(18):4009–4022, 2013.

[Wal08] D.N. Walton. *Informal Logic: A Pragmatic Approach*. Cambridge University Press, 2008.

[xxix] ¬Prop 1

[xxx] Prop 3

[Wea03] Brian Weatherson. Are you a sim? *The Philosophical Quarterly*, 53(212):425–431, 2003.

[Weh15] Dustin Wehr. Rigorous deductive argumentation for socially relevant issues. *CoRR*, abs/1502.02272, 2015.

5 Proofs (supplemental)

Notation: To cut down on some of the clutter, we drop the cardinality function symbol when it is easily inferred from the context. Specifically, whenever a finite set valued term S appears where a number is expected, it is shorthand for $|S|$.

5.1 Proof of Theorem 1

Recall the statement:

> *Let $\models_1$ denote entailment with respect to Definition 1 (models for the Patch 1 Simulation Argument). For all settings of the parameters $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$:*
>
> $$N > \mathit{LB}(d, \vec{k}), \mathit{Patch\ 1} \models_1 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3}$$

Proof. This proof builds on the proof in the appendix of [BK10]. Let $d, \vec{k}$ be arbitrary, and assume all of

$$N > \mathsf{LB}(d, \vec{k}), \quad \mathsf{Patch\ 1}, \quad \neg\mathsf{Prop\ 3}, \quad \neg\mathsf{Prop\ 2}$$

The remainder of the proof is to derive Prop 1. Define

$$R \coloneqq \frac{N}{k_3 d} - \frac{1}{d} \tag{3}$$

From ¬Prop 3, ¬Prop 2 and Patch 1, we will derive:

$$\mathsf{C}^{\overline{\mathsf{PH}}} \geqslant \mathsf{C}_{\geqslant N}\,(R - k_2) \tag{4}$$

Starting from ¬Prop 3:

$$q_3 \geqslant f_{\text{sim}}$$
$$= \frac{\sum_{c \in \mathsf{C}_{<N}} \mathsf{pop}(c)\#\mathsf{sims}(c) + \sum_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c)\#\mathsf{sims}(c)}{\sum_{c \in \mathsf{C}_{<N}} \mathsf{pop}(c)\#\mathsf{sims}(c) + \sum_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c)\#\mathsf{sims}(c) + \mathsf{avgpop}_{\geqslant N}\mathsf{C}_{\geqslant N} + \mathsf{avgpop}_{<N}\mathsf{C}_{<N}}$$

Since the fraction is in [0,1], we drop a positive term above and below:

$$\geqslant \frac{\sum\limits_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c)\#\mathsf{sims}(c)}{\sum\limits_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c)\#\mathsf{sims}(c) + \mathsf{avgpop}_{\geqslant N}\mathsf{C}_{\geqslant N} + \mathsf{avgpop}_{<N}\mathsf{C}_{<N}}$$

Again since the fraction is in [0,1], we may soundly substitute in the lower bound $\sum\limits_{c \in \mathsf{C}_{\geqslant N}} \mathsf{pop}(c)\#\mathsf{sims}(c) \geqslant N\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N}$ above and below:

$$\begin{aligned}
&\geqslant \frac{N\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N}}{N\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N} + \mathsf{avgpop}_{\geqslant N}\mathsf{C}_{\geqslant N} + \mathsf{avgpop}_{<N}\mathsf{C}_{<N}} \\
&= \frac{1}{1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}\mathsf{avgpop}_{<N}}{\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N}}\right)} \\
&\geqslant \frac{1}{1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}}{\mathsf{C}_{\geqslant N}}d\right)} && \text{by Patch 1}
\end{aligned}$$

And so

$$\begin{aligned}
& q_3 \geqslant \frac{1}{1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}}{\mathsf{C}_{\geqslant N}}d\right)} \\
\leftrightarrow \quad & \frac{1}{q_3} \leqslant 1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}}{\mathsf{C}_{\geqslant N}}d\right) \\
\leftrightarrow \quad & \frac{1 + k_3}{k_3} \leqslant 1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}}{\mathsf{C}_{\geqslant N}}d\right) && \text{defn of } k_3 \\
\leftrightarrow \quad & \left(\frac{1 + k_3}{k_3} - 1\right) - \frac{1}{N} \leqslant \frac{\mathsf{C}_{<N}d}{\mathsf{C}_{\geqslant N}N} \\
\leftrightarrow \quad & \frac{1}{k_3} - \frac{1}{N} \leqslant \frac{\mathsf{C}_{<N}d}{\mathsf{C}_{\geqslant N}N} \\
\leftrightarrow \quad & \frac{N}{k_3} - 1 \leqslant \frac{\mathsf{C}_{<N}d}{\mathsf{C}_{\geqslant N}} \\
\leftrightarrow \quad & \mathsf{C}_{<N} \geqslant \frac{\mathsf{C}_{\geqslant N}}{d}\left(\frac{N}{k_3} - 1\right) \\
\leftrightarrow \quad & \mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N} + \mathsf{C}^{\overline{\mathsf{PH}}} \geqslant \frac{\mathsf{C}_{\geqslant N}}{d}\left(\frac{N}{k_3} - 1\right) && \text{since } \mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N}, \mathsf{C}^{\overline{\mathsf{PH}}} \text{ partitions } \mathsf{C}_{<N}
\end{aligned}$$

By $\neg\mathsf{Prop}\ 2 \equiv \mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N} \leqslant k_2 \mathsf{C}_{\geqslant N}$ and the previous inequality we have

$$\begin{aligned} & & k_2 \mathsf{C}_{\geqslant N} + \mathsf{C}^{\overline{\mathsf{PH}}} &\geqslant \frac{\mathsf{C}_{\geqslant N}}{d}\left(\frac{N}{k_3} - 1\right) \\ \leftrightarrow & & \mathsf{C}^{\overline{\mathsf{PH}}} &\geqslant \mathsf{C}_{\geqslant N}\left[\frac{1}{d}\left(\frac{N}{k_3} - 1\right) - k_2\right] \\ & & &= \mathsf{C}_{\geqslant N}\left[\frac{N}{k_3 d} - \frac{1}{d} - k_2\right] \\ & & &= \mathsf{C}_{\geqslant N}\left(R - k_2\right) \end{aligned}$$

So finally, Inequality (4) is proved.

From the definition of $f_{\overline{\mathrm{PH}}}$, Inequality (4), and $\neg\mathsf{Prop}\ 2$ again, we'll derive

$$f_{\overline{\mathrm{PH}}} \geqslant \frac{R - k_2}{R + 1} \tag{5}$$

$$\begin{aligned} f_{\overline{\mathrm{PH}}} &= \frac{\mathsf{C}^{\overline{\mathsf{PH}}}}{\mathsf{C}^{\mathsf{PH}} + \mathsf{C}^{\overline{\mathsf{PH}}}} \\ &= \frac{\mathsf{C}^{\overline{\mathsf{PH}}}}{\mathsf{C}_{\geqslant N} + \mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N} + \mathsf{C}^{\overline{\mathsf{PH}}}} && \text{since } \mathsf{C}_{\geqslant N}, \mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N} \text{ partitions } \mathsf{C}^{\mathsf{PH}} \\ &\geqslant \frac{\mathsf{C}^{\overline{\mathsf{PH}}}}{\mathsf{C}_{\geqslant N} + k_2 \mathsf{C}_{\geqslant N} + \mathsf{C}^{\overline{\mathsf{PH}}}} && \text{by } \neg\mathsf{Prop}\ 2 \\ &\geqslant \frac{\mathsf{C}_{\geqslant N}(R - k_2)}{\mathsf{C}_{\geqslant N} + k_2 \mathsf{C}_{\geqslant N} + \mathsf{C}_{\geqslant N}(R - k_2)} && \text{by Inequality (4)} \\ &= \frac{R - k_2}{1 + k_2 + R - k_2} && \mathsf{C}_{\geqslant N} \text{ cancels} \\ &= \frac{R - k_2}{R + 1} && \text{Equation (3)} \end{aligned}$$

If we can prove

$$\textbf{Goal:} \qquad \frac{R - k_2}{R + 1} > 1 - q_1$$

then we're done, since then $f_{\overline{\mathrm{PH}}} \geqslant 1 - q_1$, which is equivalent to $f_{\mathrm{PH}} \leqslant q_1$, which is $\mathsf{Prop}\ 1$.

We finally use the N lower bound assumption:

$$N > \mathsf{LB}(d, \vec{k}) = d k_3 (k_1 + k_2 + k_1 k_2) + k_3$$

$N > dk_3(k_1 + k_2 + k_1k_2) + k_3$ is equivalent to

$$\begin{aligned} k_1 + k_2 + k_1k_2 &< \frac{N}{dk_3} - \frac{1}{d} \\ &= R \end{aligned}$$

Solving for k_1 in the inequality just derived, obtain:

$$k_1 < \frac{R - k_2}{1 + k_2}$$

Since by definition $k_1 = \frac{1-q_1}{q_1}$:

$$\frac{1 - q_1}{q_1} < \frac{R - k_2}{1 + k_2}$$

Solving for q_1, obtain:

$$q_1 > \frac{1}{\frac{R-k_2}{1+k_2} + 1}$$

Thus

$$\begin{aligned} 1 - q_1 &< 1 - \frac{1}{\frac{R-k_2}{1+k_2} + 1} \\ &= 1 - \frac{1}{\frac{R+1}{1+k_2}} \\ &= \frac{\frac{R+1}{1+k_2} - 1}{\frac{R+1}{1+k_2}} \\ &= \frac{\frac{R-k_2}{1+k_2}}{\frac{R+1}{1+k_2}} \\ &= \frac{R - k_2}{R + 1} \end{aligned}$$

That completes the proof. □

5.2 Proof of Corollary 1

Recall the statement:

> *Let $\models_2$ denote entailment with respect to Definition 6. For all settings of the parameters $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$:*
>
> $$N > \mathit{LB}(d, \vec{k}), \mathit{Patch\ 2} \models_2 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3'}$$

Proof. Fix a setting of the parameters. Let $\mathcal{M}$ be any 2-model that satisfies $N > \mathsf{LB}(d, \vec{k})$ and Patch 2. We construct a 1-model $\mathcal{N}$ that satisfies $N > \mathsf{LB}(d, \vec{k})$ and Patch 1, and so by Theorem 1 we get that $\mathcal{N}$ satisfies Prop 1 $\vee$ Prop 2 $\vee$ Prop 3. Lastly, we observe that this implies $\mathcal{M}$ satisfies Prop 1 $\vee$ Prop 2 $\vee$ Prop 3′.

Recall that a 2-model is just a 1-model with an additional function count^E. $\mathcal{N}$'s interpretation of every symbol of the language of 1-models *except* for pop is the same as $\mathcal{M}$'s interpretation, and the definition of $\mathcal{N}$ is completed by defining

$$(\mathsf{pop}(c))^{\mathcal{N}} = \left(\mathsf{count}^E(c)\right)^{\mathcal{M}} \text{ for every civilization } c$$

We are free to set the parameters of Theorem 1, but we have made them all the same as they are for Corollary 1, so clearly $\mathcal{N}$ satisfies $N > \mathsf{LB}(d, \vec{k})$. Also observe that

$$(\mathsf{avgpop}^E_{<N})^{\mathcal{M}} = (\mathsf{avgpop}_{<N})^{\mathcal{N}} \text{ and } (\mathsf{avgpop}^E_{\geqslant N})^{\mathcal{M}} = (\mathsf{avgpop}_{\geqslant N})^{\mathcal{N}}$$

and so $\mathcal{M}$'s satisfying Patch 2 implies $\mathcal{N}$'s satisfying Patch 1. We can now apply Theorem 1 to get that $\mathcal{N}$ satisfies Prop 1 $\vee$ Prop 2 $\vee$ Prop 3.

Observe that the meaning of Prop 1 and Prop 2 is the same for 1-models and 2-models, and by the way we defined $\mathcal{N}$ it is clear that $(\mathsf{Prop\ 1})^{\mathcal{N}} \leftrightarrow (\mathsf{Prop\ 1})^{\mathcal{M}}$ and $(\mathsf{Prop\ 2})^{\mathcal{N}} \leftrightarrow (\mathsf{Prop\ 2})^{\mathcal{M}}$. If we can show $(\mathsf{Prop\ 3'})^{\mathcal{N}} \leftrightarrow (\mathsf{Prop\ 3'})^{\mathcal{M}}$, then we're done. For that, simply note that $(\#S^E)^{\mathcal{M}} = (\#S)^{\mathcal{N}}$ and $(\#U^E)^{\mathcal{M}} = (\#U)^{\mathcal{N}}$, so $(f^E_{\text{sim}})^{\mathcal{M}} = (f_{\text{sim}})^{\mathcal{N}}$.

This model translation requires the permissibility in Definition 1 of civilizations that run at least one ancestral simulation but have cumulative pre-PH population size 0; such civilizations in $\mathcal{N}$ are produced from civilizations in $\mathcal{M}$ with no E-observers that run at least one ancestral simulation. That makes little sense according to the informal interpretation of C as a set of "civilizations". However, it is important to note that this is not a weakness. That Theorem 1 works with such models is just a mathematical fact, and we can exploit that fact to get this easy proof of Corollary 1. Alternatively, we could copy the proof of Theorem 1 and superficially modify it to get a proof of Corollary 1. □

5.3 Proofs that Theorem 1 and Corollary 1 are optimal

5.3.1 Proof of Theorem 2

Recall the statement of Theorem 2:

> *Let $\models_1$ denote entailment with respect to Definition 1 (models for the Simulation Argument). For all settings of $d \in \mathbb{N}, \vec{k} \in (\mathbb{N}^+)^3$:*
>
> $$N \geqslant \mathit{LB}(d, \vec{k}), \mathit{Patch\ 1} \not\models_1 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3}$$

Proof. We give a model (Definition 1) that satisfies $N = \mathsf{LB}(d, \vec{k})$ and Patch 1 and falsifies each of Prop 1, Prop 2, Prop 3.

We specify exactly one PH civilization[xxxi] $c_{\geqslant N}$ that does N ancestor simulations. We specify k_2 PH civilizations that do fewer than N simulations (in fact they do none), so $|\mathsf{C}_{\geqslant N}| = k_2|\mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N}|$ and Prop 1 is falsified.

Note that $|\mathsf{C}^{\mathsf{PH}}| = 1 + k_2$. We specify $k_1(1 + k_2)$ civilizations that never reach a PH state, so $|\mathsf{C}^{\overline{\mathsf{PH}}}| = k_1|\mathsf{C}^{\mathsf{PH}}|$, and Prop 2 is falsified. Note that $|\mathsf{C}_{<N}| = |\mathsf{C}^{\mathsf{PH}} \cap \mathsf{C}_{<N}| + |\mathsf{C}^{\overline{\mathsf{PH}}}| = k_2 + k_1(1 + k_2)$.

$c_{\geqslant N}$ has cumulative population size one, so $\mathsf{avgpop}_{\geqslant N} = 1$.[xxxii] For the other civilizations $\mathsf{C}_{<N}$, we specify that each has cumulative population size d, so $\mathsf{avgpop}_{<N} = d$, and Patch 1 is satisfied.

Observe that the total number of simulated observers $\#S$ is N, and the total number of non-simulated observers $\#U$ is $\mathsf{avgpop}_{\geqslant N}|\mathsf{C}_{\geqslant N}| + \mathsf{avgpop}_{<N}|\mathsf{C}_{<N}| = 1 + d(k_2 + k_1(1 + k_2))$. We'll show $\#S = k_3\#U$ so that Prop 3 is falsified, and then we're done. Indeed the reader can check that the right hand sides of the following equations are equivalent.

$$\#S = N = dk_3(k_1 + k_2 + k_1k_2) + k_3$$

$$k_3\#U = k_3(1 + d(k_2 + k_1(1 + k_2)))$$

□

[xxxi] This construction generalizes for $|\mathsf{C}_{\geqslant N}|$ equal to any positive integer.
[xxxii] The construction generalizes for $\mathsf{avgpop}_{\geqslant N}$ equal to any positive integer.

5.3.2 Proof of Corollary 2

Recall the statement of Corollary 2:

> *Let $\models_2$ denote entailment with respect to Definition 1 (models for the Simulation Argument). For all settings of $d \in \mathbb{N}, \vec{k} \in (\mathbb{N}^+)^3$:*
>
> $$N \geqslant \mathit{LB}(d, \vec{k}), \mathit{Patch\ 2} \not\models_2 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3'}$$

Proof. The proof is almost identical to that of Theorem 2. Use the same construction of a 1-model $\mathcal{M}$, and then additionally specify $\mathsf{count}^E(c) = \mathsf{pop}(c)$ for every civilization c. Note that this makes $|\mathsf{C}^{E \geqslant 1}_{\geqslant N}| = |\mathsf{C}_{\geqslant N}| = H_s(E)$ and $|\mathsf{C}^{E \geqslant 1}_{<N}| = |\mathsf{C}_{<N}| = H_n(E)$ (recall $H_n(E)$ is [BK10]'s notation for the average number of E-observers in $\mathsf{C}^{E \geqslant 1}_{<N}$ civilizations, and similarly for $H_s(E)$ and $\mathsf{C}^{E \geqslant 1}_{\geqslant N}$), in which case $\mathsf{Patch\ 2}$ and $\mathsf{Patch\ 1}$ are equivalent.

In an earlier draft of this paper, we made the fractions $\frac{|\mathsf{C}^{E \geqslant 1}_{\geqslant N}|}{|\mathsf{C}_{\geqslant N}|}$ and $\frac{|\mathsf{C}^{E \geqslant 1}_{<N}|}{|\mathsf{C}_{<N}|}$ be parameters of the argument. This makes it more tedious to prove the natural analog of Corollary 2, in which "all settings" includes setting those fractions to arbitrary rational numbers in $[0, 1]$, and letting d be any rational number greater than 0. One must construct a model with civilizations that have exactly the right ratios of E-observers to observers. It does however work out fine, after some minor adjustments to the statement of Corollary 2 (e.g. the LB term gets put inside $\lfloor \cdot \rfloor$). □