Correcting Errors in the Bostrom/Kulczycki Simulation Arguments draft, extended with proofs

R. Dustin Wehr

www.logicforprogress.org

February 26, 2017

Abstract

Both patched versions of the Bostrom/Kulczycki simulation argument contain serious objective errors, discovered while attempting to formalize them in predicate logic. The English glosses of both versions involve badly misleading meanings of vague magnitude terms, which their impressiveness benefits from. We fix the errors, prove optimal versions of the arguments, and argue that both are much less impressive than they originally appeared. Finally, we provide a guide for readers to evaluate the simulation argument for themselves, using well-justified settings of the argument parameters that have simple, accurate statements in English, which are easier to understand and critique than the statements in the original paper.

TL;DR - go straight to page 22

Contents

This note concerns the widely-known Simulation Arguments, first published by Bostrom in [Bos03], and "patched" in two different ways years later in [BK10] to correct an error. Here, we correct three further serious errors, one in the Patch 1 version and two in the Patch 2 version. We also improve the formalization, give complete proofs, and demonstrate the significance of the errors. For the corrected proofs, we show that the impressiveness of the informal English statements is largely due to an accidental equivocation pattern, which [Wal08] calls *variability of strictness of standards*, with the problematic terms used in [BK10] including "significant number of," "very likely," "extremely small," "almost certainly," "astronomically large," etc (see Sections 1.1, 2.1, and 3 in particular). 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 using a strong/large-magnitude interpretation for some assumptions, and a weak/small-magnitude interpretation for others.[i] In Section 3 we give new concise statements of the corrected simulation arguments, which should be used in preference to those in [BK10] or [Bos03].

This paper assumes familiarity with [BK10].

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

Note 1. This paper is narrowly focused, and does not attempt to provide a summary of other kinds of criticisms of the simulation argument [Wea03] [Bru08] [Lew13] [Bir13], which might be more important than this quite technical one. Also, throughout the paper, whenever possible we use via quotation the informal English glosses of [Bos03] and [BK10] for axioms, definitions and theorems, rather than introducing

[i] [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].

our own. In general, we take a completely uncritical stance on the nuanced epistemological issues involved, focusing, essentially, only on the math and logic.

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 $\mathsf{Patch\ 2}$ version), but also many mathematical symbols that have standard meanings, which would be fixed in the definition of the set of $\mathcal{L}$-structures. We 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[ii], or just 1-model, is given by:

- A finite nonempty set C, for the "human-level technological civilizations" [BK10]. We 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 PH stage.
- For each $c \in \mathsf{C}$, a natural number $\mathsf{pop}(c)$, for the cumulative pre-PH population size 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].[iii]

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

[iii] 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 we make it part of the model.

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.
- $\mathsf{C}_{<N} \coloneqq \{c \in \mathsf{C} \mid \#\mathsf{sims}(c) < N\}$ 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 **u̲**nsimulated observers.[iv]
- $\#S \coloneqq \sum_{c \in \mathsf{C}^{\mathsf{PH}}} \mathsf{pop}(c) \cdot \#\mathsf{sims}(c)$ is the total number of s̲imulated 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.[v]
- $\mathsf{avgpop}_{<N} \coloneqq \left(\sum_{c \in \mathsf{C}_{<N}} \mathsf{pop}(c) \right) / |\mathsf{C}_{<N}|$ 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 (with different notation) and definitions from [BK10].

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:

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

Prop 1 is intended to express "The human species is very likely to go extinct before

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

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

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.

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

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

$\mathsf{Prop\ 2}$ 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 ($f_{\mathrm{sim}}, q_3, \mathsf{Prop\ 3}$)**.** 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

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

$\mathsf{Prop\ 3}$ is intended to express "We are almost certainly living in a computer simulation." Formally:

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

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)**.**

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

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

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

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

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

[vi] 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.

[vii] 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.

[viii] Technically, (2) is only stronger than (1) when $N \leqslant 10^{12}$. As we note in Section 3, such a large value of N makes Prop 2 very likely, but not necessarily for an interesting reason.

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

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

However, a massively larger lower bound on N, close to one trillion (see Corollary on page 9), is needed when the erroneous (2) is replaced with (1). This is Error 1.

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 $(0, 1)$,[x] and replace them with $\mathbb{R}^+$ 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):

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

Now we can state the main results for Patch 1. Let $\vDash_1$ denote entailment for 1-models (Definition 1).[xi]

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

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

[xi] That is, if A is a sentence and Γ a set of sentences, then $\Gamma \vDash_1 A$ means every 1-model that satisfies each sentence in Γ must also satisfy A.

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

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

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

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

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:[xiii]

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

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

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

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.

[xii] It is not a mistake that the domain of $\vec{k}$ is different here. The proof of Theorem 2 (6.3.1), as it is now, uses the fact that k is an integer, whereas the proof of Theorem 1 does not. Thanks to an anonymous refereefor pointing out that this deserves explanation, since it seems like it could be an error.

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

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

> 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, we 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 we 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. Substituting in the settings $\vec{k}$ 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 $\mathsf{Patch\ 2}$ 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 $\mathsf{Patch\ 1}$ argument, from which the authors of [BK10] justify the jump

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

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 $\mathsf{Patch\ 1}$ Simulation Agument from the previous section, we will accept the qualitative assumptions of the $\mathsf{Patch\ 2}$ argument uncritically, focusing only on the mathematical aspects.

The $\mathsf{Patch\ 2}$ argument was not fleshed out in [BK10]. We do that here, and it turns out that the mathematics of the $\mathsf{Patch\ 1}$ and $\mathsf{Patch\ 2}$ arguments are practically the same. The 6.2 follows easily from Theorem 1.

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

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, $\mathsf{avgpop}^E_{\geqslant N} = \left[\sum_{c \in \mathsf{C}_{\geqslant N}} \mathsf{count}^E(c)\right] / |\mathsf{C}_{\geqslant N}|$.

[xv] 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 4).

- $\mathsf{C}_{<N}^{E\geqslant 1}$ and $\mathsf{C}_{\geqslant N}^{E\geqslant 1}$, 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>u</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 <u>s</u>imulated observers.

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

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

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

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

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

The quantitative part[xvi] of the informal definition of 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 $H_s(E)$.

[xvi] See footnote xxiii on page 19 for a third item (iii) about not having "other information that enables us to tell that we are not in a simulation." Error 3 relates to that.

Unfortunately, that does not quite suffice; this is 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. 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*

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

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.

The reader should convince themself intuitively at this point that 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 this 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.[xvii] That is the approach we take in the proof of Corollary 1 (6.2).

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

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

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

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

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

Note 3. The Patch 1 argument is a special case of the Patch 2 argument; just take E to be "I am a member of my civilization."

We 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 (6.2) and Corollary 2 6.3.2) easier.

2.1 Effect on the argument

The English glosses of [BK10]'s Patch 2 result suffers from a problem similar to that of their statement 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

[xvii] 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 3), 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. We give further explanation of why such reasoning is beyond reproach in the appendix.

1.1 we showed that 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 12 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:

“X is vastly greater than Y” means

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

Thus, *just* to get the meaning of “X is vastly greater than Y” to be the trivial $X > Y$, we need the meaning of “significant number” to be nearly 1 million. We 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 2argument 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, Error 3, and Guide to Evaluating the Simulation Argument for Yourself

The reader may wonder why we 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, we wanted to be sure that we were treating not just the published form of Bostrom/Kulczycki's arguments fairly, but also the ideas behind that form.

Thus, we did not settle for merely demonstrating absurd consequences of the errors in the arguments, as we did in Sections 1.1 and 2.1, since such a thing could in principle be fixed; we 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 use, in this section, settings of the parameters that require knowing the exact form of LB.

<u>**Observation**</u>: 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 (and Error 3), so a reader who prefers the Patch 1 argument can stop just before then, and use $E =$ "I am a member of my civilization" to reduce to the Patch 1 argument, instead of the E we 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, we 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 $> k_1$ 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 $> k_2$ times greater* *than the number that run at least N ancestor simulations.*

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

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

Let $\vDash_2$ denote entailment with respect to Definition 6. Then

$$N > dk_3(k_1 + k_2 + k_1k_2) + k_3 \vDash_2 \textit{Prop 1} \vee \textit{Prop 2} \vee \textit{Prop 3} \vee \textit{Prop 4}$$

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

$$N \geqslant dk_3(k_1 + k_2 + k_1k_2) + k_3 \nvDash_2 \textit{Prop 1} \vee \textit{Prop 2} \vee \textit{Prop 3} \vee \textit{Prop 4}$$

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

When $k_1k_2k_3$[xx] 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 Simulation Arguments exploits variability of strictness of standards, as mentioned in the introduction.

[xviii] Formerly ¬Patch 2.

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

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

3.1 The special case of Theorem 3 that you should pay attention to, and Error 3

Let A-Civilizations abbreviate "advanced human-like civilizations".

Step 1

Set $k_3 = 1$, so that Prop 3 becomes "There are more simulated E-observers than non-simulated E-observers." Then, **if you accept [BK10]'s interpretation of the proposition using the "bland indifference principle"**, you get

> Prop 3: You are probably simulated (more likely than not).

Convince yourself that this simplification does not undercut the strength of the Simulation Argument. In particular, convince yourself that we should make Prop 2 and Prop 4 as unlikely as possible (since they're less interesting than Prop 1 and Prop 3) *subject to the constraint* that Prop 3 remains both profound and easy to internalize. Minimizing k_3 makes Prop 2 and Prop 4 as unlikely as possible.

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[xxi]. [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 $k_2 = 99$, **if you accept [BK10]'s reasoning that we should, roughly, treat our civilization as a random sample from the A-Civilizations**, then you get

[xxi] 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.

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

Step 3 (and Error 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. To see this, simply note that the main novelty of Bostrom's Simulation Argument is the use of the assumption

$$P(p \text{ is simulated} \,|\, p\text{'s original}^{\text{xxii}}\text{civilization does} \geqslant N \text{ ancestor simulations}) \geqslant \frac{N}{N+1}$$

which is must closer to 1 than any reasonable prior $P(p \text{ is simulated})$. The authors of [BK10] miss that point with their example E[xxiii](this is **Error 3**).

Thus, "My computer age birth rank is X" is not sufficient. You must at least strengthen it to:

[xxii] Unsimulated

[xxiii] They are, however, aware that care must be taken in defining E, as evidenced by item (iii) (page 4) of their sketch of Patch 2, which we reprint here:

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

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.

> 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 world, the militaries of the two nations with the largest nuclear arsenals are commanded by thugs, ⟨something about climate change⟩, etc...

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

> Prop 4 version (a): The fraction of E-observer-having civilizations among A-Civilizations that run fewer than N ancestor simulations is more than d times greater than the fraction of E-observer-having civilizations among A-Civilizations that run at least N ancestor simulations.

Optional Step 4

Set $d = 1$, as suggested by one of the authors of [BK10] in private communication. Then Prop 4 further simplifies to:

> Prop 4 version (b): The frequency of E-observer-having civilizations among our-technology-or-better civilizations that run fewer than N ancestor simulations is greater than the frequency of E-observer-having civilizations among our-technology-or-better civilizations that run at least N ancestor simulations.

[BK10] argues that the previous simplified version of Prop 4 is false, which would take us back to a tripartite disjunction, but that is too generous in light of Error 3 (subsection Step 3 (and Error 3) on page 19).

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.

I believe the following is essentially the most compelling single parameter version of [BK10]'s repaired Simulation Argument possible. If you wish to consider two parameter versions, which I believe benefits the Simulation Argument, I recommend stopping the simplifications before Optional Step 4.

We give the simplified English statement of the theorem for $k_2 = 1$ in the next section.

3.1.1 Example 4-disjunct 1-Parameter Simplified English Version

Let k_2 be any positive real number. Recall A-Civilizations abbreviates "advanced human-like civilizations". See Step 3 (and Error 3) for the meaning of "E-observer". Then at least one of the following is true:

- Prop 1: There's at least a 99% chance that the human race is destroyed before it reaches the posthuman stage.
- Prop 2: There are some A-Civilizations that reach the posthuman stage, but among *those* there are at least k_2 times as many that run fewer than $N = 100(k_2 + 1)$ ancestor simulations than there are that run at least N ancestor simulations.
- Prop 3: You are probably simulated (more likely than not).
- Prop 4: The fraction of E-observer-having civilizations among A-Civilizations that run fewer than N ancestor simulations $<$ than the fraction of E-observer-having civilizations among A-Civilizations that run at least N ancestor simulations.

We find the result is easier to understand as an implication, putting the negations of the more-technical Prop 2 and Prop 4 on the left hand side and the clearly-profound Prop 1 $\vee$ Prop 3 on the right hand side, as follows.

3.1.2 Example Implication form, 1-Parameter Simplified English Version

Let k_2 be any positive real number. Recall A-Civilizations abbreviates “advanced human-like civilizations”. See Step 3 (and Error 3) for the meaning of “E-observer”. Assume that at least one of the A-Civilizations reach the posthuman stage, since otherwise Prop 1 is trivially true.

If

> ¬Prop 2: Among the A-Civilizations that eventually reach the posthuman stage, there are less than k_2 times as many that run fewer than $N = 100(k_2+1)$ ancestor simulations than there are that run at least N ancestor simulations,

and

> ¬Prop 4: The fraction of E-observer-having civilizations among A-Civilizations that run fewer than N ancestor simulations $\geqslant$ the fraction of E-observer-having civilizations among A-Civilizations that run at least N ancestor simulations.

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 (more likely than not)”).

Observe that the dependence of N on k_2 makes for a very difficult task of setting k_2 to maximize the acceptability of ¬Prop 2 in the antecedent. Making k_2 large (say, 99 as in [BK10]) seems at first like a good strategy of making the antecedent true, but then you notice that doing so raises N, and so shifts more of the eventually-posthuman civilizations into the category of running fewer than N ancestor simulations.

3.1.3 Example Implication form, 0-Parameter Simplified English Version

Here we do a final simplification, fixing the parameter k_2 to 1.

Recall A-Civilizations abbreviates "advanced human-like civilizations". See Step 3 (and Error 3) for the meaning of "E-observer". Assume that at least one of the A-Civilizations reach the posthuman stage, since otherwise Prop 1 is trivially true.

If

> $\neg$Prop 2: Among the A-Civilizations that eventually reach the posthuman stage, there are fewer that run < 200 ancestor simulations than there are that run $\geqslant 200$ ancestor simulations,

and

> $\neg$Prop 4: The fraction of E-observer-having civilizations among A-Civilizations that run < 200 ancestor simulations is greater than the fraction of E-observer-having civilizations among A-Civilizations that run $\geqslant 200$ ancestor simulations.

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 (more likely than not)").

3.1.4 Why I suspect Prop 2 is true (making the implication trivial)

I will make this quick. First, in such posthuman civilizations that are so profoundly advanced that they can build a simulation as realistic as what we perceive as real life, wealthy individuals would simply have too little to gain to put in the expense of fully programming something so complex, and then running it N times. We must be very careful not to succumb to the issue of variability of strictness of standards (mentioned in the introduction) when it comes to evaluating the plausibility of this. In particular, the more feasible it is for an institution in an A-Civilization to run an ultra-realistic ancestor simulation (and some argue it's not possible at all), the less interesting it will be to do so.

Second, I believe that in a civilization capable of surviving long enough to run such simulations, most people will believe that running such realistic ancestor simulations is profoundly unethical, since it involves, for example, prolonged torturing of individuals, who they would consider sentient, in ways that would astonish any of us who has not done a brief review of the ghastly history of torture. I, for example, would ardently advocate for and participate in pre-emptive military action against any institution who was planning such a thing. If that view was popular, ancestor simulations would need to be carried out in secret. Moreover, for Prop 2 to be false, it would need to be more common than not that in such posthuman civilizations, the psychopaths responsible for secretly perfecting such ancestor simulations[xxiv] run, on average, at least 200 such simulations before getting bored (for $k_2 = 1$).

4 Other pernicious vagueness

Proponents of the simulation argument, when faced with the criticism of the plausibility of a civilization running stupendously expensive ancestor simulations, will respond with the possibility of software tricks to reduce the *energy* cost, e.g.

- classical approximations of quantum physics whenever possible (since there is no good reason to have confidence that quantum computing will *ever* be a reality).
- editing of the simulation when bugs arise.

[xxiv] I say perfecting since, in order to count in the argument, an ancestor simulation needs to reach our point in technological advancement before crashing or veering off course.

- starting the simulation from a "recent" but complex state, e.g. the beginning of life on Earth rather than the big bang.

Those are not complete responses, since those tricks impose a much greater cost of software development and during-simulation maintenance. For example, the first idea presumes that our descendants will be able to specify when quantum-classical differences affect the simulation in a way that humans would notice, and when they don't. Even if our descendants have superintelligence-powered program generators, there is no shortcut for software specification. The second requires manual modification of the state, and a software model that supports that (and we have no good reason to believe that neural nets will ever support that). The third requires specifying the initial state; again, there is no shortcut for writing such a complex specification.

In general, do not let the question of the cost of running N ultra realistic ancestor simulations, which is dominated by energy, distract you from the cost of running *one* ultra realistic ancestor simulation, which is dominated by the effort of intelligent agents, and constrained by the nature of reality.

5 Conclusion

We corrected three remaining serious errors in two "patched" versions of the Simulation Argument, and analyzed their significance carefully in the language of mathematical logic. We found that, although the corrected arguments are sound, their meaning is highly dependent on the settings of their (interdependent) parameters, and we do not believe the parameters can be set in a way that makes the Simulation Argument nearly as impressive as it appeared in [BK10], where the example parameter settings turned out to rely on inconsistent interpretation of vague magnitude terms.

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.

[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.

6 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 $|S|$.

6.1 Proof of Theorem 1

Recall the statement:

> *Let $\vDash_1$ denote entailment with respect to Definition 1 (models for the Patch 1 Simulation Argument). For all settings of the parameters $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$:*
>
> $$N > \mathit{LB}(d, \vec{k}), \mathit{Patch\ 1} \vDash_1 \mathit{Prop\ 1} \vee \mathit{Prop\ 2} \vee \mathit{Prop\ 3}$$

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:

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

Starting from $\neg\mathsf{Prop\ 3}$:

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

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

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

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

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

$$\geqslant \frac{N\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N}}{N\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N} + \mathsf{avgpop}_{\geqslant N}\mathsf{C}_{\geqslant N} + \mathsf{avgpop}_{<N}\mathsf{C}_{<N}}$$

$$= \frac{1}{1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}\mathsf{avgpop}_{<N}}{\mathsf{C}_{\geqslant N}\mathsf{avgpop}_{\geqslant N}}\right)}$$

$$\geqslant \frac{1}{1 + \frac{1}{N}\left(1 + \frac{\mathsf{C}_{<N}}{\mathsf{C}_{\geqslant N}}d\right)} \qquad \text{by } \mathsf{Patch\ 1}$$

And so

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

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

$$
\begin{aligned}
& & k_2\mathsf{C}_{\geqslant N}+\mathsf{C}^{\overline{\mathsf{PH}}} &\geqslant \frac{\mathsf{C}_{\geqslant N}}{d}\left(\frac{N}{k_3}-1\right) \\
&\leftrightarrow & \mathsf{C}^{\overline{\mathsf{PH}}} &\geqslant \mathsf{C}_{\geqslant N}\left[\frac{1}{d}\left(\frac{N}{k_3}-1\right)-k_2\right] \\
& & &= \mathsf{C}_{\geqslant N}\left[\frac{N}{k_3d}-\frac{1}{d}-k_2\right] \\
& & &= \mathsf{C}_{\geqslant N}\left(R-k_2\right)
\end{aligned}
$$

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:

$$N > \mathsf{LB}(d, \vec{k}) = dk_3(k_1 + k_2 + k_1 k_2) + k_3$$

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

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

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. □

6.2 Proof of Corollary 1

Recall the statement:

> *Let $\vDash_2$ denote entailment with respect to Definition 6. For all settings of the parameters $d \in \mathbb{N}, \vec{k} \in (\mathbb{R}^+)^3$:*
>
> $$N > \mathsf{LB}(d, \vec{k}), \mathsf{Patch\ 2} \vDash_2 \mathsf{Prop\ 1} \vee \mathsf{Prop\ 2} \vee \mathsf{Prop\ 3'}$$

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

Observe that the meaning of $\mathsf{Prop\ 1}$ and $\mathsf{Prop\ 2}$ 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 $(f^{E}_{\mathrm{sim}})^{\mathcal{M}} = (f_{\mathrm{sim}})^{\mathcal{N}}$.

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. □

6.3 Proofs that Theorem 1 and Corollary 1 are optimal

6.3.1 Proof of Theorem 2

Recall the statement of Theorem 2:

> *Let $\vDash_1$ denote entailment with respect to Definition 1 (models for the Simulation Argument). For all settings of $d \in \mathbb{N}, \vec{k} \in (\mathbb{N}^+)^3$:*
>
> $$N \geqslant \textsf{\textit{LB}}(d, \vec{k}), \textsf{\textit{Patch 1}} \nvDash_1 \textsf{\textit{Prop 1}} \vee \textsf{\textit{Prop 2}} \vee \textsf{\textit{Prop 3}}$$

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

We specify exactly one PH civilization[xxv] $c_{\geqslant N}$ 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 $\mathsf{Prop\ 1}$ 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 $\mathsf{Prop\ 2}$ 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 $\mathsf{avgpop}_{\geqslant N} = 1$.[xxvi] For the other civilizations $\mathsf{C}_{<N}$, we specify that each has cumulative population size d, so $\mathsf{avgpop}_{<N} = d$, and $\mathsf{Patch\ 1}$ 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 $\mathsf{Prop\ 3}$ is falsified, and then we're done. Indeed the reader can check that the right hand sides of the following equations are equivalent.

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

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

[xxvi] The construction generalizes for $\mathsf{avgpop}_{\geqslant N}$ equal to any positive integer.

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

□

6.3.2 Proof of Corollary 2

Recall the statement of Corollary 2:

> *Let $\vDash_2$ denote entailment with respect to Definition 1 (models for the Simulation Argument). For all settings of $d \in \mathbb{N}, \vec{k} \in (\mathbb{N}^+)^3$:*
>
> $$N \geqslant \mathsf{LB}(d, \vec{k}), \mathsf{Patch\ 2} \nvDash_2 \mathsf{Prop\ 1} \vee \mathsf{Prop\ 2} \vee \mathsf{Prop\ 3'}$$

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 $\lfloor \cdot \rfloor$). □