How the Simulation Argument Dampens Future Fanaticism

Brian Tomasik
Center on Long-Term Risk
brian.tomasik@longtermrisk.org

June 2016

Abstract

Some effective altruists assume that most of the expected impact of our actions comes from how we influence the very long-term future of Earth-originating intelligence over the coming ∼billions of years. According to this view, helping humans and animals in the short term matters, but it mainly only matters via effects on far-future outcomes.

There are a number of heuristic reasons to be skeptical of the view that the far future astronomically dominates the short term. This piece zooms in on what I see as perhaps the strongest concrete (rather than heuristic) argument why short-term impacts may matter a lot more than is naively assumed. In particular, there's a non-trivial chance that most of the copies of ourselves are instantiated in relatively short-lived simulations run by superintelligent civilizations, and if so, when we act to help others in the short run, our good deeds are duplicated many times over. Notably, this reasoning dramatically upshifts the relative importance of short-term helping *even if* there's only a small chance that Nick Bostrom's basic simulation argument is correct.

My thesis doesn't prove that short-term helping is more important than targeting the far future, and indeed, a plausible rough calculation suggests that targeting the far future is still several orders of magnitude more important. But my argument does leave open uncertainty regarding the short-term-vs.-far-future question and highlights the value of further research on this matter.

Contents

1 Epigraph

> The question is whether one can get more value from controlling structures that — in an astronomical-sized universe — are likely to exist many times, than from an extremely small probability of controlling the whole thing.
>
> – steven0461

2 Introduction

One of the ideas that's well accepted within the effective-altruism community but rare in the larger world is the immense importance of the far-future effects of our actions. Of course, many environmentalists are concerned about the future of Earth, and people in past generations have started projects that would not finish in their lifetimes. But it's rare for in-the-trenches altruists, rather than just science-fiction authors and cosmologists, to consider the effects of their actions on sentient beings that will exist billions of years from now.

Future focus is extremely important, but it can at times be exaggerated. It's sometimes thought that the far future is so important that the short-term effects of our actions on the welfare of organisms alive today are negligible by comparison, *except* for instrumental reasons insofar as short-term actions influence far-future outcomes. I call this "far-future fanaticism". I probably believed something along these lines from ∼2006 to ∼2013.

However, like with almost everything else in life, the complete picture is more complicated. We should be extremely suspicious of any simple argument which claims that one action is, say, 10^30 times more important than another action, e.g., that influencing the far future is 10^30 times more important than influencing the near term. Maybe that's true, but reality is often complex, and extraordinary claims of that type should not be accepted hastily. This is one of several reasons we should maintain modesty about whether working to influence the far future is vastly better than working to improve the wellbeing of organisms in the nearer term.

3 Anti-mugging approaches

Dylan Matthews, like many others, has expressed skepticism about far-future fanaticism on the grounds that it smells of Pascal's mugging. I think far-future fanaticism is a pretty mild form of (mugger-less) Pascal's mugging, since the future fanatic's claim is vastly more probable *a priori* than the Pascal-mugger's claim. Still, Pascal's mugging comes in degrees, and lessons from one instance should transfer to others.[1]

3.1 Hansonian leverage penalty

The most popular resolution of Pascal's mugging on the original thread was that by Robin Hanson: "People have been talking about assuming that states with many people hurt have a low (prior) probability. It might be more promising to assume that states with many people hurt have a low *correlation* with what any random person claims to be able to effect."

[1] Eliezer Yudkowsky would probably dislike my characterization of far-future focus as a mild form of Pascal's mugging:

> the phrase "Pascal's Mugging" got completely bastardized to refer to an emotional feeling of being mugged that some people apparently get when a high-stakes charitable proposition is presented to them, regardless of whether it's supposed to have a low probability. This is enough to make me regret having ever invented the term "Pascal's Mugging" in the first place [...].

Of course, influencing the far future does have a lower probability of success than influencing the near term. The difference in probabilities is just relatively small (plausibly within a few orders of magnitude).

ArisKatsaris generalized Hanson's idea to "The Law of Visible Impact": "Penalize the prior probability of hypotheses which argue for the existence of high impact events whose consequences nonetheless remain unobserved."

Eliezer Yudkowsky called this a "leverage penalty". However, he goes on to show how a leverage penalty against the possibility of helping, say, a googleplex people can lead you to disbelieve scenarios where you could have huge impact, no matter how much evidence you have, which seems possibly wrong.

3.2 Simulation argument

In this piece, I don't rely on a general Hansonian leverage penalty. Rather, I use the simulation argument, which resembles the Hansonian leverage penalty in its effects, but it does so organically rather than in a forced way.

Yudkowsky says: "Conceptually, the Hansonian leverage penalty doesn't interact much with the Simulation Hypothesis (SH) at all." However, the two ideas act similarly and have a historical connection. Indeed, Yudkowsky discussed something like the simulation-argument solution to Pascal's mugging after hearing Hanson's idea:

> Yes, if you've got 3↑↑↑↑3 people running around they can't *all* have sole control over each other's existence. So in a scenario where lots and lots of people exist, one has to penalize by *a proportional factor* the probability that any one person's binary decision can solely control the whole bunch.
>
> Even if the Matrix-claimant says that the 3↑↑↑↑3 minds created will be unlike you, with information that tells them they're powerless, if you're in a generalized scenario where anyone has and uses that kind of power, the vast majority of mind-instantiations are in leaves rather than roots.

The way I understand Yudkowsky's point is that if the universe is big enough to contain 3↑↑↑↑3 people, then for every person who's being mugged by a genuine mugger with control over 3↑↑↑↑3 people, there are probably astronomical numbers of other people who are confronting lying muggers, pranks, hallucinations, dreams, and so on. So across the multiverse, almost all people who get Pascal-mugged can't actually save 3↑↑↑↑3 people, and in fact, the number of people who get fake Pascal-mugged is proportional to 3↑↑↑↑3. Hence, the probability of *actually* being able to help N people is roughly k/N for some constant k, so the expected value of giving in to the mugging remains finite regardless of how big N is.

However, this same kind of reasoning also works for Yudkowsky's "Pascal's Muggle" scenario in which a Matrix Lord opens "up a fiery portal in the sky" to convince a person that the Matrix Lord is telling the truth about a deal to save a googleplex lives for $5. But given that there's a huge amount of computing power in the Matrix Lord's universe, for every one Matrix Lord who lets a single person determine the fate of a googleplex people, there may be tons of Matrix Lords just faking it (whether for the lulz, to test the simulation software, or for some other reason). So the expected number of copies of a person facing a lying Matrix Lord should be proportional to a googleplex, and hence, the probability penalty that the Hansonian prior would have suggested seems roughly vindicated. Yudkowsky makes a similar point:

> when it comes to improbability on the order of 1/3↑↑↑3, the prior improbability *is* inescapable - your sensory experiences *can't* possibly be that unique - which is assumed to be appropriate because almost-everyone who ever believes they'll be in a position to help 3↑↑↑3 people *will in fact* be hallucinating. Boltzmann brains should be much more common than people in a unique position to affect 3↑↑↑3 others, at least if the causal graphs are finite.

3.3 Reliance on observers?

ArisKatsaris complains that Hanson's principle "seems to treat the concept of 'person' as ontologically fundamental", the way that other instances of Nick Bostrom-style anthropic reasoning do (Bostrom, 2010). But, with the simulation-argument approach, you can avoid this problem by just talking about exact copies of yourself, where a "copy" means "a physical structure whose high-level decision-making algorithms exactly mirror your own, such that what you decide to do, it also decides to do". A copy needn't (and in general doesn't) share your full environment, just your current sensory inputs and behavioral outputs for some (possibly short) length of time. Then Yudkowsky's argument is that almost all copies of you are confronting fake or imagined muggers.

3.4 Application to future fanaticism

We can apply the simulation anti-mugging argument to future fanaticism. Rather than being the sole person out of $3\uparrow\uparrow\uparrow\uparrow3$ people to control the actions of the mugger, we on Earth in the coming centuries are, perhaps, the sole tens of billions of people to control the far-future of Earth-originating intelligence, which might involve $\sim 10^{52}$ people, to use the Bostrom estimate quoted in Matthews's article. For every one biological human on the real Earth, there may be tons of simulated humans on simulated Earths, so most of our copies probably "are in leaves rather than roots", to use Yudkowsky's terminology.

Even if Earth-originating intelligence specifically doesn't run ancestor simulations, other civilizations may run simulations, such as when studying the origin of life on various planets, and we might be in some of those simulations. This is similar to how, even though a real Pascal-mugger might specify that all of the $3\uparrow\uparrow\uparrow\uparrow3$ people that *she* will create will never think they're being Pascal-mugged, in the multiverse at large, there should be lots more people in various other circumstances who *are* fake Pascal-mugged.

Yudkowsky acknowledges the simulation possibility and its implications for future fanaticism:

> If we *don't* take everything at face value, then there might be such things as ancestor simulations, and it might be that your experience of looking around the room is something that happens in 10^{20} ancestor simulations for every time that it happens in 'base level' reality. In this case your probable leverage on the future is diluted (though it may be large even post-dilution).

If we think of ourselves as all our copies rather than a particular cluster of cells or transistors, then the simulation hypothesis doesn't decrease our probable leverage but actually increases it, especially the leverage from short-term actions, as is discussed below.

4 Simulation argument upshifts the relative importance of short-term helping

I first began thinking about this topic due to a post by Pablo Stafforini:

> if you think there is a chance that posthumanity will run ancestor simulations [...], the prospect of human extinction is much less serious than you thought it was.

Since I'm a negative utilitarian, I would probably prefer for space not to be colonized, but Stafforini's point also has relevance for efforts to reduce the badness of the far future, not just efforts to prevent human extinction.

Robin Hanson makes a similar point:

> if not many simulations last through all of human history, then the chance that your world will end soon is higher than it would be if you were not living in a simulation.

> So all else equal you should care less about the future of yourself and of humanity, and live more for today. This remains true even if you are highly uncertain of exactly how long the typical simulation lasts.

One response is to bite the simulation bullet and just focus on scenarios where we are in fact in basement-level reality, since if we are, we can still have a huge impact: "Michael Vassar - if you think you are Napoleon, and everyone that thinks this way is in a mental institution, you should still act like Napoleon, because if you are, your actions matter a lot."

A second response is to realize that actions focused on helping in the short term may be relatively more important than the future fanatic thought. Most simulations are probably short-lived, because one can run lots of short-lived simulations with the same computing resources as it takes to run a single long-lived simulation. Hedonic Treader: "Generally speaking, it seems that if you have evidence that your reality may be more short-lived than you thought, this is a good reason to favor the near future over the far future."

5 How much does the simulation argument reduce future fanaticism?

Note: This section is a more detailed version of an argument written here. Readers may find that presentation of the calculations simpler to understand.

This section presents a simplified framework for estimating the relative importance of short-term vs. far-future actions in light of the simulation argument. An example of an action targeted for short-term impact is changing ecosystems on Earth in order to reduce wild-animal suffering, such as by converting lawns to gravel. An example of a far-future-focused action is spreading the idea that it's wrong to run detailed simulations of ecosystems (whether for reasons of science, entertainment, or deep ecology) because of the wild-animal suffering they would contain. Of course, both of these actions affect both the short term and the far future, but for purposes of this analysis, I'll pretend that gravel lawns only prevent bugs from suffering in the short run, while anti-nature-simulation meme-spreading only helps prevent bugs from suffering in the long run. I'm trying to focus just on the targeted impact time horizon, but of course, in reality, even if the future fanatic is right, every short-term action has far-future implications, so no charity is 10^30 times more important than another one.

So, a non-trivial fraction of all technological civilizations, both human and alien, colonize the cosmos and create capacious computing clusters. I'll assume that most of the suffering of the far future will be created by the computations that an advanced civilization would run. Rather than measuring computational capacity in FLOPS or some other conventional performance metric, I'll measure computations by how much sentience they contain in the form of the agents and subroutines that are being computed, with the unit of measurement being what I'll call a "sent". I define "sentience" as "morally relevant complexity of mental life". I compute the moral value (or disvalue) for an agent experiencing an emotion as

> moral value = (sentience of the agent) · (how intense the agent would judge the emotion to be relative to evolutionarily/physiologically typical emotions for that agent) · (duration of the experience).

For example, if a human has sentience of 1 sent and a fly has sentience of 0.01 sents, then even if a fly experiences a somewhat more damaging event relative to its utility function, that event may get less moral weight.

Using units of sentience will help make later calculations easier. I'll define 1 sent-year as the amount of complexity-weighted experience of one life-year of a typical biological human. That is, consider the sentience over time experienced in a year by the median

biological human on Earth right now. Then, a computational process that has 46 times this much subjective experience has 46 sent-years of computation.[2] Computations with a higher density of sentience may have more sents even if they have fewer FLOPS.

Suppose there's a large but finite number C of civilizations that are about to colonize space. (If one insists that the universe is infinite, one can restrict the analysis to some huge but finite subset of the universe, to keep infinities from destroying math.) On average, these civilizations will run computations whose sentience is equivalent to that of N human-years, i.e., a computing capacity of N sent-years. So these civilizations collectively create the equivalent of $C \cdot N$ sent-years.

Some of these minds may be created by agents who want to feel intense emotions by immersing (copies of) themselves in experience-machines or virtual worlds. Also, we have much greater control over the experiences of a programmed digital agent than we do over present-day biological creatures.[3] These factors suggest that influencing a life-year experienced by a future human might be many times more altruistically important than influencing a life-year experienced by a present-day human. The future, simulated human might have much higher intensity of experience per unit time, and we may have much greater control over the quality of his experience. Let the multiplicative factor T represent how much more important it is to influence a unit of sentience by the average future digital agent than a present-day biological one for these reasons. T will be in units of moral (dis)value per sent-year. If one thinks that a significant fraction of post-human simulations will be run for reasons of wireheading or intrinsically valuing intense experiences, then T may be much higher than 1, while if one thinks that most simulations would be run for purposes of scientific / historical discovery, then T would be closer to 1. T also counts the intensity and controllability of non-simulation subjective experiences. If a lot of the subjective experience in the far future comes from low-level subroutines that have fairly non-intense experiences, then T might be closer to 1.

Suppose that the amount of sentience on Earth in the near term (say, the next century or two) is some amount E sent-years. And suppose that some fraction f_E of this sentience takes the form of human minds, with the rest being animals, other life forms, computers, and so on.

Some far-future simulations may contain just one richly computed mind in an otherwise superficial world. I'll call these "solipsist simulations". Many other simulations may contain several simulated people interacting but in a very limited area and for a short time. I'll neologize the adjective "solipsish" to refer to these simulations, since they're not quite solipist, but because they have so few people, they're solipsist-ish. Robin Hanson paints the following picture of a solipsish simulation:

> Consider, for example, a computer simulation of a party at the turn of the millennium created to allow a particular future guest to participate. This simulation might be planned to last only one night, and at the start be limited to the people in the party building, and perhaps a few people visible from that building. If the future guest decided to leave the party and wander the city, the simulated people at the party might be erased, to be replaced by simulated people that populate the street where the partygoer walks.

In contrast, a non-solipsish simulation is one in which most or all of the people and animals who seem to exist on Earth are actually being simulated to a non-trivial level of detail. (Inanimate matter and outer space may still be simulated with low levels of richness.)

[2] 1 sent-year for simulated humans will probably take place in much less than 1 sidereal year, assuming simulations have high clock speeds.

[3] This is particularly true for increasing happiness, where in biological creatures we face the hedonic treadmill. It's less true in the case of a negative utilitarian reducing suffering by decreasing population size, since preventing an individual from existing completely eliminates its suffering, whether it's biological or digital.

Let f_N be the fraction of computations run by advanced civilizations that are non-solipsish simulations of beings who think they're humans on Earth, where computations are measured in sent-years, i.e., f_N = (sent-years of all non-solipsish sims who think they're humans on Earth)/(sent-years of all computations that are run in total). And let f_C be the fraction of the C civilizations who actually started out as biological humans on Earth (rather than biological aliens).

5.1 Calculation using Bostrom-style anthropics and causal decision theory

I and most MIRI researchers have moved on from Bostrom-style anthropic reasoning, but Bostrom anthropics remains well known in the scholarly literature and is useful in many applications, so I'll first explore the implications of the simulation argument in this framework. In particular, I'll use the self-sampling assumption with the reference class of "humans who think they're on pre-colonization Earth". The total number of such humans is a combination of those who *actually are* biological organisms on Earth:

(number of real Earths) · (human sent-years per real Earth) =
$(C \cdot f_C) \cdot (E \cdot f_E)$

and those in simulations who *think* they're on Earth:

(number of advanced-civilization computations) · (fraction comprised of non-solipsish humans who think they're on Earth) = $C \cdot N \cdot f_N$.

Note that Bostrom's strong self-sampling assumption samples randomly from observer-moments, rather than from sent-years, but assuming all humans have basically the same sentience, then sampling from sent-years should give basically the same result as sampling from observer-moments.

Horn #3 of Bostrom's 2003a simulation-argument trilemma can be seen by noting that as long as N/E is extremely large (reject horn #1) and f_N / ($f_C \cdot f_E$) is not correspondingly extremely tiny (reject horn #2), the ratio of simulated to biological humans will be very large:

$$\frac{\text{non-solipsish simulated human sent-years}}{\text{biological human sent-years}} = \frac{C \cdot N \cdot f_N}{C \cdot f_C \cdot E \cdot f_E} = \frac{\frac{N}{E} \cdot f_N}{f_C \cdot f_E}$$

If you are sampled randomly from all (non-solipsish) simulated + biological human sent-years, the probability that you are a biological human, P_b , is

$$P_b = \frac{\text{biological human sent-years}}{(\text{simulated hum an sent-years}) + (\text{biological human sent-years})} = \frac{C \cdot f_C \cdot E \cdot f_E}{(C \cdot N \cdot f_N) + (C \cdot f_C \cdot E \cdot f_E)} = \frac{f_C \cdot E \cdot f_E}{(N \cdot f_N) + (f_C \cdot E \cdot f_E)}$$

If we are biological humans, then we're in a position to influence all of the N expected sent-years of computation that lie in our future, which will have, on average, higher intensity and controllability by a factor of T units of moral value per sent-year. On the other hand, it's much harder to reliably influence the far future, because there are so many unknowns and so many intervening steps in the causal chain between what we do now and what happens centuries or gigayears from now. Let D be a discount representing how much harder it is to actually end up helping a being in the far future than in the

near term, due to both uncertainty and the muted effects of our actions now on what happens later on.

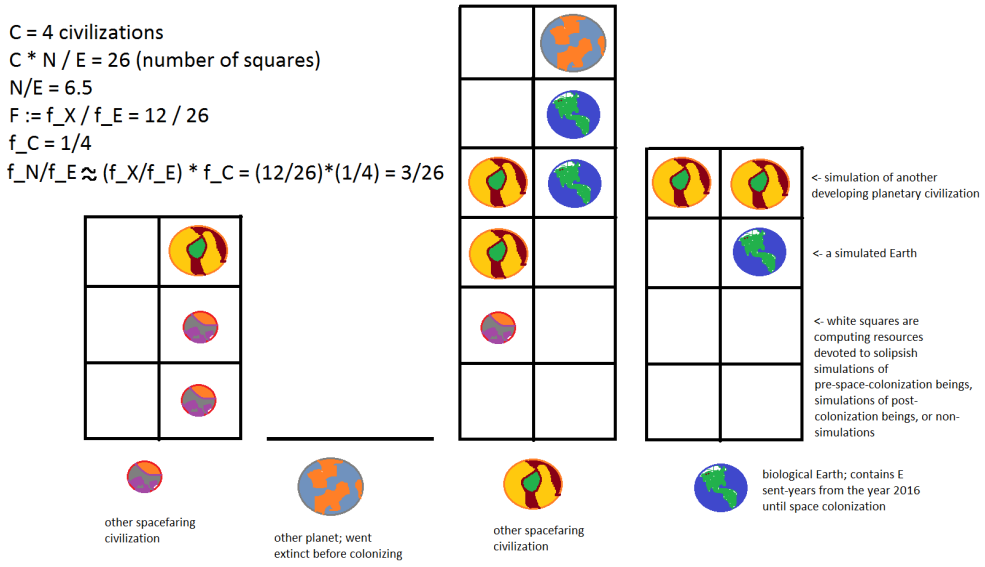
If we are biological humans, then targeting the far future can affect N expected sent-years with intensity multiple of T, but with discount D, for an expected impact proportional to $N \cdot T \cdot D$.[4] On the other hand, if we target the short term, we can help the sentience currently on Earth, with an impact proportional to E.[5]

However, actions targeting the far future only matter if there is a far future. In most simulations, the future doesn't extend very far, because simulating a long post-human civilization would be extremely computationally expensive. For example, emulating a planet-sized computer in a simulation would probably require at least a planet-sized computer to run the simulation. As an approximation, let's suppose that actions targeting far-future impact only matter if we're biological humans on an actual Earth. Then the expected impact of far-future actions is proportional to $P_b \cdot N \cdot T \cdot D$. Let's call this quantity "L" for "long-term impact". In contrast, actions targeting the short term make a difference whether we're simulated or not, as long as the simulation runs for at least a few decades and includes most animals on Earth. So the expected impact of short-term-focused actions is just E. Let's call our expected impact for short-term actions S.

The ratio of these two quantities is $L / S = P_b \cdot N \cdot T \cdot D / E$.

5.1.1 A simple example

The following picture shows a cartoon example of the framework I'm using here. I haven't yet defined all the variables that you see in the upper left corner, but they'll be explained soon.



Note that $N = 6.5 \cdot E$ and $f_N = (3/26) \cdot f_E$. By inspecting the picture, we can see that P_b should be $1/4$, since there's one real Earth and three simulated versions. As hoped, our formula for P_b verifies this:

[4]The units in the product $N \cdot T \cdot D$ are (number of sent-years) · (moral value of helping a given sent-year)·(probability discount on actually helping any given sent-year).

[5]The units here are (E sent-years) · (1 unit of moral value per sent-year). The intensity factor here is 1 unit of moral value per sent-year, since the intensity factor T for long-term helping was defined relative to the intensity factor for short-term helping. There's no probability discount here, because the long-term discount D was defined as the probability discount for long-term helping *relative* to short-term helping.

$$\begin{aligned} P_b &= \frac{f_C \cdot E \cdot f_E}{(N \cdot f_N) + (f_C \cdot E \cdot f_E)} \\ &= \frac{\frac{1}{4} \cdot E \cdot f_E}{(6.5 \cdot E \cdot \frac{3}{26} \cdot f_E) + (\frac{1}{4} \cdot E \cdot f_E)} \\ &= \frac{\frac{1}{4}}{(6.5 \cdot \frac{3}{26}) + \frac{1}{4}} \\ &= \frac{1}{4} \end{aligned}$$

And

$$\frac{L}{S} = P_b \cdot N \cdot T \cdot \frac{D}{E} = \frac{1}{4} \cdot 6.5 \cdot T \cdot D = 1.6 \cdot T \cdot D$$

Note that in the actual picture, Earth has 8 squares of far-future computation ahead of it, but N/E is only 6.5. That's because N/E is an average across civilizations, including some that go extinct before colonizing space. But an average like this seems appropriate for our situation, because we don't know *ex ante* whether humanity will go extinct or how big humanity's computing resources will be compared with those of other civilizations.

5.2 Calculation based on all your copies

Now I'll redo the calculation using a framework that doesn't rely on the self-sampling assumption. Rather, it takes inspiration from anthropic decision theory (Armstrong, 2011). You should think of yourself as all your copies at once. Rather than thinking that you're a single one of your copies that might be biological or might be simulated, you should think of yourself as *both* biological *and* simulated, since your choices affect both biological and simulated copies of you. The interesting question is what the ratio is of simulated to biological copies of you.

When there are more total copies of Earth (whether biological or simulated), there will be more copies of you. In particular, suppose that some constant fraction f_y of all non-solipsish human sent-years (whether biological or simulated) are copies of you. This should generally be roughly the case, because a non-solipsish simulation of Earth-in-the-year-2016 should have ~ 7 billion humans in it, one of which is you.

Then the expected number of biological copies (actually, copy life-years) of you will be $f_y \cdot C \cdot f_C \cdot E \cdot f_E$, and the expected number of simulated copy life-years will be $f_y \cdot C \cdot N \cdot f_N$.[6]

Now suppose you take an action to improve the far future. All of your copies, both simulated and biological, take this action, although it only ends up mattering for the biological copies, since only they have a very long-term future. For each biological copy, the expected value of the action is proportional to $N \cdot T \cdot D$, as discussed in the previous subsection. So the total value of having all your copies take the far-future-targeting

[6]Note that these expressions assume that the sentience of all your copies is the same, since they assume a constant ratio f_y that converts from sent-years of general humans to life-years for one of your copies. However, we might care a bit less about copies of ourselves that are simulated in lower-resolution simulations (e.g., simulations that only represent a crude neuronal level of detail rather than a sub-neuronal level of detail, assuming the high-level behavior of the brain is the same in both cases). If the sentience of everyone else in a low-resolution simulation is lower to the same degree that your copy's sentience is lower, then the sent-years that the copy in the low-res simulation will be able to help will be correspondingly lower. In such a case, it would be ok for the calculations in this piece to count ourselves as having only, say, 1/3 of a copy in a low-res simulation whose sent-years are 1/3 as much as normal, as long as the amount of helping the copy could do would also be only 1/3 as much on average. That's because this piece assumes that the amount of short-term helping we can do is proportional to the number of copies we have. In other words, we can think of a copy as "a unit of helping power", with lower-resolution instances of ourselves being less than one full copy because they have less helping power.

action is proportional to

L = (number of biological copies of you) · (expected value per copy) = (f_y · C · f_C · E · f_E) · (N · T · D)

In contrast, consider taking an action to help in the short run. This helps whether you're biological or non-solipsishly simulated. The expected value of the action for each copy is proportional to E, so the total value across all copies is proportional to

S = (number of biological + non-solipsish simulated copies of you) · (expected value per copy) = (f_y · C · f_C · E · f_E + f_y · C · N · f_N) · E

Then we have

$$\frac{L}{S} = \frac{(f_y \cdot C \cdot f_C \cdot E \cdot f_E) \cdot (N \cdot T \cdot D)}{(f_y \cdot C \cdot f_C \cdot E \cdot f_E + f_y \cdot C \cdot N \cdot f_N) \cdot E}$$

Interestingly, this exactly equals P_b · N · T · D / E, the same ratio of far-future vs. short-term expected values that we calculated using the self-sampling assumption.

5.3 Simplifying L/S

Simplifying the L/S expression above:

$$\begin{aligned}\frac{L}{S} &= \frac{[N \cdot T \cdot \frac{D}{E}] \cdot (f_C \cdot E \cdot f_E)}{(f_C \cdot E \cdot f_E) + (N \cdot f_N)} \\ &= \frac{T \cdot D \cdot f_C}{f_C \cdot E/N + f_N/f_E}\end{aligned}$$

Note that this ratio is strictly less than T · D · f_C / (f_N/f_E), which is a quantity that doesn't depend on N. Hence, we can't make L/S arbitrarily big just by making N arbitrarily big.

Let f_X be the average fraction of superintelligent computations devoted to non-solipsishly simulating the development of any almost-space-colonizing civilization that actually exists in biological form, not just humans on Earth. f_N is the fraction of computations devoted to simulating humans on Earth in particular. If we make the simplifying assumption that the fraction of simulations of humans on Earth run by the collection of all superintelligences will be proportional to the fraction of humans out of all civilizations in the universe, then $f_N = f_X \cdot f_C$. This would be true if

- all civilizations run simulations of all other civilizations in proportion to their numerosity
- only human descendants (not aliens) run simulations of only humans on Earth (not of aliens) and have a typical amount of computing power devoted to such simulations, or
- various combinations in between these extremes is true.

Making this assumption, we have

$$\frac{L}{S} = \frac{T \cdot D \cdot f_C}{f_C \cdot \frac{E}{N} + f_X \cdot \frac{f_C}{f_E}} = \frac{T \cdot D}{\frac{E}{N} + \frac{f_X}{f_E}}$$

Non-solipsish simulations of the dominant intelligences on almost-space-colonizing planets also include the (terrestrial or extraterrestrial) wild animals on the same planets. Assuming that the ratio of (dominant-intelligence biological sent-years)/(all biological

sent-years) on the typical almost-space-colonizing planet is approximately f_E , then f_X/f_E would approximately equal the fraction of all computational sent-years spent non-solipsishly simulating almost-space-colonizing ancestral planets (both the most intelligent and also less intelligent creatures on those planets). I'll call this fraction simply F. Then

$$\frac{L}{S} = \frac{T \cdot D}{\frac{E}{N} + F}$$

Visualized using the picture from before, f_N /f_E is the fraction of squares with Earths in them, and F is the fraction of squares with any planet in them.

Everyone agrees that E/N is very small, perhaps less than 10^{-30} or something, because the far future could contain astronomical amounts of sentience (Bostrom, 2003b). If F is not nearly as small (and I would guess that it's not), then we can approximate L/S as T · D / F.

5.4 Plugging in parameter values

Now that we have an expression for L/S, we'd like to know whether it's vastly greater than 1 (in which case the far-future fanatics are right), vastly less than 1 (in which case we should plausibly help beings in the short run), or somewhere in the ballpark of 1 (in which case the issue isn't clear and needs more investigation). To do this, we need to plug in some parameters.

Here, I'll plug in point estimates of T, D, and F, but doing this doesn't account for uncertainty in their values. Formally, we should take the full expected value of L with respect to the probability distributions of T and D, and divide it by the full expected value of S with respect to the probability distribution for F. I'm avoiding that because it's complicated to make up complete probability distributions for these variables, but I'm trying to set my point estimates closer to the variables' expected values than to their median values. Our median estimates of T, D, and F are probably fairly different from the expected values, since extreme values may dominate the expected-value calculations. For this reason, I've generally set the parameter point estimates higher than I actually think is reasonable as a median estimate. And of course, your own estimates may be pretty different.

5.4.1 D = 10^{-3}

This is because (a) it's harder to know if a given action now will actually have a good impact in the long term than it is to know that a given action will have a good impact in the short term and (b) while a single altruist in the developed world can exert more than a ∼1/(7 billion) influence on all the sentience on Earth right now (such as by changing the amount of wilderness that exists), a single person may exert less than that amount of influence on the sentience of the far future, because there will be generations after us who may have different values and may override our decisions.

In particular, for point (a), I'm assuming a ∼0.1 probability discount, because, for example, while it's not implausible to be 75% confident that a certain action will reduce short-run wild-animal populations (with a 25% chance of increasing them, giving a probability discount of $75\% - 25\% = 50\%$), on many far-future questions, my confidence of making a positive rather than negative impact is more like 53% (for a probability discount of $53\% - 47\% = 6\%$, which is about 10 times smaller than 50%).

For point (b), I'm using a ∼0.01 probability discount because there may be generations ahead of us before the emergence of artificial general intelligence (AGI), and even once AGI arrives, it's not clear that the values of previous humans will translate into the values of the AGI, nor that the AGI will accomplish goal preservation without further mutation of those values. Maybe goal preservation is very difficult to implement or is

strategically disfavored by a self-improvement race against aliens, so that the changes to the values and trajectory of AGI we work toward now will be overridden thousands or millions of years later. (Non-negative utilitarians who consider preventing human extinction to be important may not discount as much here because preventing extinction doesn't have the same risk of goal/institutional/societal drift as trying to change the future's values or general trajectory does.)

5.4.2 $\mathbf{T = 10^4}$

Some simulations run by superintelligences will probably have extremely intense emotions, but many (especially those run for scientific accuracy) will not. Even if only an expected 0.01% of the far future's sent-years consist of simulations that are 10^8 times as intense per sent than average experiences on Earth, we would still have $T \sim 10^4$.

5.4.3 $\mathbf{T = 10^{-6}}$

It's very unclear how many simulations of almost-space-colonizing planets superintelligences would run. The fraction of all computing resources spent on this might be close to 100% or might be below 10^{-15} . It's hard to predict resource allocation by advanced civilizations. But I set this parameter based on assuming that $\sim 10^{-4}$ of sent-years will go toward ancestor simulations *of some sort* (this is probably too high, but it's biased upward in expectation, since, e.g., maybe there's a 0.05% chance that post-humans devote 20% of sent-years to ancestor simulations), and only 1% of those simulations will be of the almost-space-colonizing period (since there might also be many simulations of the origin of life, prehistory, and the early years after a planet's "singularity"). If we think that simulations contain more sentience per petaflop of computation than do other number-crunching calculations, then 10^{-4} of sent-years devoted to ancestor simulations of some kind may mean less than 10^{-4} of all raw petaflops devoted to such simulations.

5.4.4 Calculation using point estimates

Using these inputs, we have

$$\frac{L}{S} \sim T \cdot \frac{D}{F} = 10^4 \cdot \frac{10^{-3}}{10^{-6}} = 10^7$$

This happens to be bigger than 1, which suggests that targeting the far future is still $\sim$ 10 million times better than targeting the short term. But this calculation could have come out as less than 1 using other possible inputs. Combined with general model uncertainty, it seems premature to conclude that far-future-focused actions dominate short-term helping. It's likely that the far future will still dominate after more thorough analysis, but by much less than a naive future fanatic would have thought.

6 Objections

6.1 Doesn't this assume that the simulation hypothesis is 99.999999% likely to be true?

No. My argument works as long as one maintains only at least a modest probability (say, at least 1% or 0.01%) that the simulation hypothesis is correct.

If one entirely rejects the possibility of simulations of almost-space-colonizing civilizations, then F = 0. In that case,

$$\frac{L}{S} = \frac{T \cdot D}{\frac{E}{N} + F} = T \cdot D \cdot \frac{N}{E},$$

which would be astronomically large because N/E is astronomically large. So if we were certain that $F = 0$ (or even that F was merely on the order of E/N in size), then we would return to future fanaticism. But we're not certain of this, and our impact doesn't become irrelevant if $F > 0$. Indeed, the more simulations of us there are, the more impact we have by short-term-targeting actions!

Let's call a situation where F is on the order of E/N in size or smaller the "F_{tiny}" possibility, and the situation where F is much bigger than E/N the "$F_{moderate}$" possibility. The expected value of $S, E[S]$, is

$$E[S|F_{tiny}] \cdot P(F_{tiny}) + E[S|F_{moderate}] \cdot P(F_{moderate})$$

and similarly for E[L]. While it's true that E[S | tiny_F] is quite small, because in that case we don't have many copies in simulations, E[S | moderate_F] is bigger. Indeed,

$$\begin{aligned}\frac{E[L]}{E[S]} &= \frac{E[L]}{E[S|F_{tiny}] \cdot P(F_{tiny}) + E[S|F_{moderate}] \cdot P(F_{moderate})} \\ &\leq \frac{E[L]}{E[S|F_{moderate}] \cdot P(F_{moderate})} \\ &\sim \frac{E[L|F_{moderate}]}{E[S|F_{moderate}] \cdot P(F_{moderate})},\end{aligned}$$

where the last line assumes that L isn't drastically affected by the value of F. This last expression is very roughly like (L/S) / P(moderate_F), where L/S is computed by plugging in some moderate value of F like I did with my sample numbers above. So unless you think P(moderate_F) is extremely small, the overall E[L]/E[S] ratio won't change dramatically upon considering the possibility of no simulations.

I've heard the following defense made of future fanaticism against simulations:

1. Due to model uncertainty, the probability that I'm not in a simulation is non-vanishing.
2. Therefore, the probability that I can have astronomical impact by far-future efforts is non-vanishing.
3. But I can't have astronomical impact by short-term efforts.
4. So the far future dominates in expectation.

This reply might work if you only consider yourself to be a single one of your copies. But if you correctly realize that your cognitive algorithms determine the choices of all of your copies jointly, then it's no longer true that short-term-focused efforts don't have astronomical impacts, because there are, in expectation, astronomical numbers of simulated copies of you in which your good deeds are replicated.

6.2 What if almost all civilizations go extinct before space colonization?

This objection suggests that horn #1 of Bostrom's trilemma may be true. If almost all technological civilizations fail to colonize space – whether because they destroy themselves or because space colonization proves infeasible for some reason – this would indeed dramatically reduce the number of advanced computations that get run, i.e., N would be quite small.

I find this possibility unlikely, since it seems hard to imagine why basically all civilizations would destroy themselves, given that humanity appears like it has a decent shot at colonizing space. Maybe it's more likely that there are physical/technological limitations on massive space colonization.

But if so, then the far future probably matters a lot less than it seems, either because

humanity will go extinct before long or because, even if humans do survive, they won't create astronomical numbers of digital minds. Both of these possibilities downplay future fanaticism. Maybe the far future could matter quite a bit more than the present if humanity survives another ~ 100 million years on Earth, but without artificial general intelligence and robust goal preservation, it seems much harder to ensure that what we do now will have a reliable impact for millions of years to come (except in a few domains, like maybe affecting CO_2 emissions).

6.3 What if most of the simulations are long-lived?

In the previous argument, I assumed that copies of us that live in simulations don't have far futures ahead of them because their simulations are likely to end within decades, centuries, or millennia. But what if the simulations are very long-lived?

It seems unlikely a simulation could be as long-lived as the basement-level civilization, since it's plausible that simulating X amount of computations in the simulation requires more than X basement computations. But we could still imagine, for example, 2 simulations that are each 1/5 as big as the basement reality. Then aiming for far-future impact in those simulations would still be pretty important, since our copies in the simulations would affect 2 far futures each 1/5 as long as the basement's far future.

Note that my argument's formalism already accounts for this possibility. F is the fraction of far-future computations that simulate almost-space-colonizing planets. Most of the far future is not at the almost-space-colonizing stage but at the space-colonizing stage, so most computations simulating far-future outcomes don't count as part of F. For example, suppose that there's a basement reality that simulates 2 far-future simulations that each run 1/5 as long as the basement universe runs. Suppose that pre-space-colonizing planets occupy only 10^{-20} of all sentience in each of those simulations. Ignoring the non-simulation computations also being run, that means $F = 10^{-20}$, which is very close to 0. So the objection that the simulations that are run might be very long can be reduced to the objection that F might be extremely close to zero, which I discussed previously. The generic reply is that it seems unreasonable to be confident that F is so close to zero, and it's quite plausible that F is much bigger (e.g., 10^{-10}, 10^{-5}, or something like that). If F is bigger, short-term impact is replicated more often and so matters relatively more.

I would expect some distribution of lengths of simulations, perhaps following a power law. If we look at the distribution of lengths of threads/processes that run on present-day computers, or how long companies survive, or almost anything similar, we tend to find a lot of short-lived things and a few long-lived things. I would expect simulations to be similar. It seems unreasonable to think that across all superintelligences in the multiverse, few short-lived simulations are run and the majority of simulations are long.

Another consideration is that if the simulators know the initial conditions they want to test with the simulation, then allowing the simulation to run longer might mean that it increasingly diverges from reality as time goes on and errors accumulate.

Also, if there are long-lived simulations, they might themselves run simulations, and then we might have short-lived copies within those nested simulations. As the number of levels of simulation nesting goes up, the length (and/or computational complexity) of the nested simulations must go down, because less and less computing power is available (just like less and less space is available for the innermost matryoshka dolls).

If the far future was simulated and the number and/or complexity of nested simulations wasn't progressively reduced as the level of nesting increased, then running simulations beyond the point when simulations became feasible would require an explosion of computing power (Jenkins, 2006):

> The creators of the simulation would likely not continue it past the point in history when the technology to create and run these simulations on a widespread basis was first developed. [...] Another reason is to avoid stacking of simulations, i.e.

> simulations within simulations, which would inevitably at some point overload the base machine on which all of the simulations are running, thereby causing all of the worlds to disappear. This is illustrated by the fact that, as Lloyd (2006) of MIT has noted in his recent book, *Programming the Universe*, if every single elementary particle in the real universe were devoted to quantum computation, it would be able to perform 10^{122} operations per second on 10^{92} bits of information. In a stacked simulation scenario, where 10^6 simulations are progressively stacked, after only 16 generations, the number of simulations would exceed by a factor of 10^4 the total number of bits of information available for computation in the real universe.

The period when a civilization is almost ready to colonize space seems particularly interesting for simulators to explore, since it crucially affects how the far future unfolds. So it would make sense that there would be more simulations of the period around now than there would be of the future 1 million years from now, and many of the simulations of the 21st century would be relatively short.

Beyond these qualitative arguments, we can make a quantitative argument as to why the far future within simulations shouldn't dominate: A civilization with N sent-years of computing power in its far future can't produce more than N sent-years of simulated far-future sentience, even if it only ran simulations and had no simulation overhead (i.e., a single planet-sized simulated computer could be simulated with only a single planet-sized real computer). More likely, a civilization with N sent-years of computing power would only run like N/100 sent-years of simulated far-future sentience, or something like that, since probably it would also want to compute things besides simulations. So what's at stake with influencing the "real" far future is probably much bigger than what's a stake influencing the simulated far future. (Of course, simulated far futures could be bigger if we exist in the simulations of aliens, not just our own civilization. But unless we in particular are extremely popular simulation targets, which seems unlikely *a priori*, then in general, across the multiverse, the total simulated far futures that we control should be less than the total real far futures that we control.) Of course, a similar point applies to simulations of short-term futures: The total sent-years in all short-term futures that we control is very likely less than the total sent-years in the far futures we control (assuming we have copies both in simulations and in basement realities). The argument as to why short-term helping might potentially beat long-term helping comes from our greater ability to affect the short term and know that we're making a positive rather than negative short-term impact. Without the D probability penalty for far-future actions, it would be clear that $L > S$ within my framework.

6.4 What if the basement universe has unlimited computing power?

What if the basement universe has unbounded computing power and thus has no limitations on how long simulations can be? And what if simulations run extremely quickly, so there's no reason not to run a whole simulated universe from the big bang until the stars die out? Even then, it's not clear to me that we wouldn't get mostly short-lived simulations, especially if they're being run for reasons of intrinsic value. For every one long-lived simulation, there might be millions or quadrillions of short-lived ones.

However, one could make the argument that if the basement-level simulators are only interested in science, then rather than running short simulations (except when testing their simulation software), they might just run a bunch of long simulations and then look at whatever part of a long simulation is of interest at any given time. Indeed, they might run all possible histories of universes with our laws of physics, and once that complete collection was available to them, they wouldn't need to run any more simulations of universes with our physical laws. Needless to say, this possibility is extremely speculative. Maybe one could argue that it's also extremely important because if this scenario is true, then there are astronomical numbers of copies of us. But

there are all kinds of random scenarios in which one can raise the stakes in order to try to make some obscure possibility dominate. That is, after all, the point of the original Pascal's-mugging thought experiment. In contrast, I don't consider the simulation-based argument I'm making in this piece to be a strong instance of Pascal's mugging, because it actually seems reasonably likely that advanced civilizations will run lots of simulations of people on Earth.

In any case, even if it's true that the basement universe has unbounded computing resources and has run simulations of all possible histories of our universe, this doesn't escape my argument. The simulations run by the basement would be long-lived, yes. But those simulations would plausibly contain nested simulations, since the advanced civilizations within those simulations would plausibly want to run their own simulations. Hence, most of our copies would live in the nested simulations (i.e., simulations within simulations), and the argument in this piece would go through like before. The basement simulators would be merely like deist gods who set our universe in motion and then let it run on its own indefinitely.

6.5 Our simulated copies can still impact the far future by helping our simulators

Even if a copy of you lives in a short-lived simulation, it might have a causal impact well beyond the simulation. Many simulations may be run for reasons of scientific discovery, and by learning things in our world, we might inform our simulators of those things, thereby having a massive impact.

I find this a weak argument for several reasons.

1. If the simulators wanted to learn things about the universe in general, it would probably be more successful for them to use artificial general intelligences to do so rather than creating fake worlds filled with primates, only a fraction of whom do scientific research.
2. If we can help our simulators just by showing them how civilizations develop, that's fine, but then it's not clear that we should take any particular actions one way or another based on this possibility.
3. If we are only one out of tons of simulations, the impact of our particular information for the simulators is small. (Compare to the value of a single survey response out of a 5000-person survey.)
4. It's not clear if we want to help our simulators, since they might have values antithetical to our own.

6.6 What if simulations aren't conscious?

I'm quite confident that I would care about simulated humans. If you don't think you would, then you're also less likely to care about the far future in general, since in many far-future scenarios, especially those that contain the most sentient beings, most intelligence is digital (or, at least, non-biological; it could be analog-computed).

If you think it's a factual rather than a moral question whether simulations are conscious, then you should maintain some not-too-small probability that simulations are conscious and downweight the impact your copies would have in simulations accordingly. As long as your probability of simulations being conscious is not tiny, this shouldn't change the analysis too much.

If you have moral uncertainty about whether simulations matter, the two-envelopes problem comes to haunt you. But it's plausible that the faction of your moral parliament that cares about simulations should get some influence over how you choose to act.

6.7 The simulation argument is weird

In a post defending the huge importance of the far future, steven0461 anticipates the argument discussed in this piece:

> the idea that we're living in an ancestor simulation. This would imply astronomical waste was illusory: after all, if a substantial fraction of astronomical resources were dedicated toward such simulations, each of them would be able to determine only a small part of what happened to the resources. This would limit returns. It would be interesting to see more analysis of optimal philanthropy given that we're in a simulation, but it doesn't seem as if one would want to predicate one's case on that hypothesis.

But I think we should include simulation considerations as a strong component of the overall analysis. Sure, they're weird, but so is the idea that we can somewhat reliably influence the Virgo-Supercluster-sized computations of a posthuman superintelligence, which is the framework that the more persuasive forms of future fanaticism rely on.

6.8 Simulated people matter less due to a bigger Kolmogorov penalty

This objection is abstruse but has been mentioned to me once. Some have proposed weighing the moral value of an agent in proportion to the Kolmogorov complexity of locating that agent within the multiverse. For example, it's plausibly easier to locate a biological human on Earth than it is to locate any particular copy of that human in a massive array of post-human simulations. The biological human might be specified as "the 10,481,284,089th human born[7] since the year that humans call AD 0, on the planet that started post-human civilization", while the simulated version of that human might be "on planet #5,381,320,108, in compartment #82,201, in simulation #861, the 10,481,284,089th human born since the year that the simulated humans call AD 0". (These are just handwavy illustrations of the point. The actual descriptions would need vastly greater precision. And it's not completely obvious that some of the ideas I wrote with text could be specified compactly.) The shortest program that could locate the simulated person is, presumably, longer than the shortest program that could locate the biological person, so the simulated person (and, probably, the other beings in his simulated world) get less moral weight. Hence, the astronomical value of short-term helping due to the correlated behavior of all of that person's copies is lower than it seems.

However, a view that gives generally lower moral weight to future beings in this way should also give lower moral weight to the other kinds of sentient creatures that may inhabit the far future, especially those that are not distinctive enough to be located easily. So the importance of influencing *the far future* is also dampened by this moral perspective. It's not obvious and would require some detailed calculation to assess how this location-penalty approach affects the relative importance of short-term vs. far-future helping.

6.9 Many copies of a brain don't matter much more than one copy

earthwormchuck163: "I'm not really sure that I care about duplicates that much."[8] Applied to the simulation hypothesis, this suggest that if there are many approximate

[7] Assuming that we can specify in a simple way a unique index for any given human birth ignores complications with abortions, stillbirths, twins, whether a birth happens when the child begins or ends its exit from the birth canal, etc. For basically simultaneous births on opposite sides of the planet, the relativity of simultaneity might also become relevant.

[8] earthwormchuck163 later changed his/her mind on this point.

copies of you helping other Earthlings across many simulations, since you and the helped Earthlings have roughly the same brain states in the different simulations, those brain states might not matter a lot more than a single such brain state. In that case, your ability to help tons of copies in simulations via short-term-focused actions would be less important. In contrast, the far future may contain a large variety of minds, so even though, within the N expected sent-years that you can influence by targeting the far future, there will be plenty of duplicates, there will be fewer duplicates than in the minds you could help by targeting the short term.

My main response is that I find it wrong to consider many copies of a brain not much more important than a single brain. This just seems intuitive to me, but it's reinforced by Bostrom (2006):

> if the universe is indeed infinite then on our current best physical theories all possible human brain-states would, with probability one, be instantiated somewhere, independently of what we do. But we should surely reject the view that it follows from this that all ethics that is concerned with the experiential consequences of our actions is void because we cannot cause pain, pleasure, or indeed any experiences at all.

Another reply is to observe that whether a brain counts as a duplicate is a matter of opinion. If I run a given piece of code on my laptop here, and you run it on your laptop on the other side of the world, are the two instances of the software duplicates? Yes in the sense that the high-level logical behavior is the same. No in the sense that they're running on different chunks of physics, at different spatiotemporal locations, in the proximity of different physical objects, etc. Minds have no non-arbitrary boundaries, and the "extended mind" of the software program, including the laptop on which it's running and the user running it, is not identical in the two cases.

Finally, it's plausible that most simulations would have low-level differences between them. It's unlikely that simulations run by two different superintelligent civilizations will be exactly the same down to the level of every simulated neuron or physical object. Rather, I conjecture that there would be lots of random variation in the exact details of the simulation, but assuming your brain is somewhat robust to variations in whether one random neuron fires or not at various times, then several slightly different variations of a simulation can have the same high-level input-output behavior and thus can all be copies of "you" for decision-theoretic purposes.

Of course, perhaps the view that "many copies don't count much more than one copy" would also say that *near* copies also don't count much more than one copy. If this is your view despite the problem that any given experience happens somewhere in the multiverse (or even just in a big enough *finite* universe, with arbitrarily high probability), then perhaps the simulation argument in this essay will not be very convincing.

6.10 If we're simulated, then reducing suffering by preventing existence frees up more computing resources

This is an important and worrying consideration. For example, suppose you aim to prevent wild-animal suffering by reducing habitat and thereby decreasing wildlife populations. If the simulation includes models of the neurons of all animals but doesn't simulate inanimate matter in much detail, then by reducing wildlife numbers, we would save computing resources, which the simulators could use for other things. Worryingly, this might allow simulators to run more total simulations of Earth-like planets, most of the neurons on which are found in invertebrates who have short lives and potentially painful deaths.

If reducing wildlife by 10% allowed simulators to run 10% more total Earth simulations, then habitat reduction would sadly not reduce much suffering. [9] But if a nontriv-

[9] Habitat reduction might still reduce a tiny amount of suffering because even though the total

ial portion of the computing power of Earth simulations is devoted to not-very-sentient processes like weather, an X% reduction in wild-animal populations reduces the computational cost of the whole simulation by less than X%. Also, especially if the simulations are being run for reasons of science rather than intrinsic value, the simulators may only need to run so many simulations for their purposes, and our making the simulations cheaper wouldn't necessarily cause the simulators to run more.[10] The simulators might use those computing resources for other purposes. Assuming those other purposes would, on average, contain less suffering than exists in wilderness simulations, then reducing habitat could still be pretty valuable.

One might ask: If $T > 1$, then won't the non-Earth-simulation computations that can be run in greater numbers due to saving on habitat computations have a *greater* density of suffering, not less, than the habitat computations had? Not necessarily, because T gives the intensity of emotions per sent-year. But many of the computations that an advanced civilization would run might not contain much sentience.[11] So the intensity of emotions per petaflop-year of non-Earth-simulation computation, rather than per sent-year, might be lower than T. Nonetheless, we should worry that this might not be true, in which case reducing habitat and thereby freeing up computing resources for our simulators would be net bad (at least for negative utilitarians; for classical utilitarians, replacing ecosystems that contain net suffering with other computations that may contain net happiness may be win-win).

amount of computation being done would be the same in the two scenarios, if habitat is smaller, then a bigger fraction of computations are devoted to humans, who have better lives than wild animals. For example, suppose that if we don't reduce wild-animal habitats, there will be some number Y of simulations with a ratio of 10,000 wild-animal sent-years per human sent-year in them. And suppose that if we do reduce wild-animal habitats (by, say, an absurdly high amount: 90%), then there will be 1000 wild-animal sent-years for every 1 human sent-year. If the total sent-years of computing power devoted to such simulations is constant, then the new number of simulations, Z, will be such that

$$Y \cdot (10,000 + 1) = Z \cdot (1000 + 1),$$

i.e., $Z = 9.991 \cdot Y$. And the new amount of wild-animal suffering will be only $Z \cdot 1000 = 9.991 \cdot Y \cdot 1000 = 9,991 \cdot Y$ sent-years, rather than $10,000 \cdot Y$.

[10] Or maybe the simulators would run more cheaper simulations but not enough more to totally negate the effect of having less habitat. Picture a demand curve for simulations, where the "price" is the cost to run a single simulation. If most of a simulation's computations are devoted to the sentient parts of wilderness (rather than to not-very-sentient physical processes like weather), then decreasing wilderness by X% should decrease the cost per simulation by about X%. If demand is inelastic, then the quantity demanded (i.e., number of simulations run) won't increase as much as the per-simulation cost decreased. Suppose that price decreases by $100 * f_p$ percent, and quantity demanded increases by $100 \cdot f_q$ percent. Since demand is inelastic (i.e., elasticity is < 1),

$$\begin{aligned}
\left| \frac{\text{percent change in quantity demanded}}{\text{percent change in price}} \right| &< 1 \\
\left| \frac{100 \cdot f_q}{-100 \cdot f_p} \right| &< 1 \\
|-1| \cdot \left| \frac{f_q}{f_p} \right| &< 1 \\
\frac{f_q}{f_p} &< 1,
\end{aligned}$$

where the last line follows because f_q and f_p are both positive numbers. Finally, note that total suffering is basically (*cost per simulation*) $\cdot$ (*number of simulations*), and the new value of this product is

$$\begin{aligned}
&old_cost_per_simulation \cdot (1 - f_p) \cdot old_number_of_simulations \cdot (1 + f_q) \\
&= old_cost_per_simulation \cdot old_number_of_simulations \cdot (1 + f_q - f_p - f_p \cdot f_q),
\end{aligned}$$

which is a decrease if $f_q < f_p$. QED.

[11] That said, as Carl Shulman pointed out to me, a non-trivial fraction of wildlife simulations on Earth may also have very little sentience – e.g., the bodies of animals, weather, fires, ocean currents, etc.

7 Copies that aren't both biological and simulated simultaneously

So far I've been assuming that if there are many copies of us in simulations, there are also a few copies of us in basement reality as well at various points in the multiverse. However, it's also possible that we're in a simulation that doesn't have a mirror image in basement-level reality. For instance, maybe the laws of physics in our simulated world are different from the basement's laws of physics, and there's no other non-simulated universe in the multiverse that shares our simulated laws of physics. Maybe our world contains miracles that the simulators have introduced. And so on. Insofar as there are scenarios in which we have copies in simulations but *not* in the basement (except for extremely rare Boltzmann-brain-type copies that may exist in some basement worlds, or extremely low-measure universes in the multiverse where specific miracles are hard-coded into the basement-level laws of physics), this amplifies the value of short-term actions, since we would be able to influence our many simulated copies but wouldn't have much of any basement copies who could affect the far future.

On the flip side, it's possible that basically all our copies are in basement-level reality and don't have exact simulated counterparts. One example of why this might be would be if it's just too hard to simulate a full person and her entire world in enough detail for the person's choices in the simulation to mirror those of the biological version. For example, maybe computationally intractable quantum effects prove crucial to the high-level dynamics of a human brain, and these are too expensive to mirror in silico.[12] The more plausible we find this scenario, the less important short-term actions look. But as we've seen, unless this scenario has probability very close to 1, the ambiguity between whether it's better to focus on the short term or long term remains unresolved.

Even if all simulations were dramatically different from all basement civilizations, as long as some of the simulated creatures thought they were in the basement, the simulation argument would still take effect. If most almost-space-colonizing organisms that exist are in simulations, then it's most likely that whatever algorithm your brain is running is one of those simulations rather than in a basement universe.

I'm still a bit confused about how to do anthropic reasoning when, due to limited introspection and bounded rationality, you're not sure which algorithm you are among several possible algorithms that exist in different places. But a naive approach would seem to be to apportion even odds among all algorithms that you might be that you can't distinguish among.

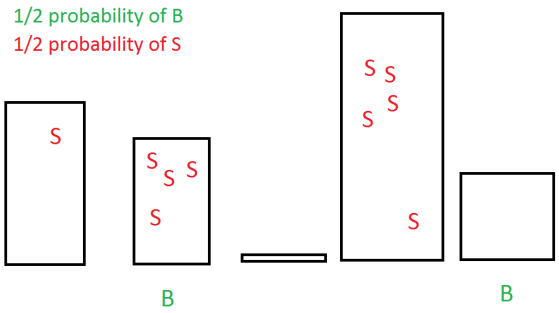
For example, suppose there are only two types of algorithms that you might be: (1)

[12] Of course, simulations needn't just use digital computation. If, for some reason, the quantum effects of biological neurons are essential for the algorithms that human brains perform, and these algorithms can't be simulated on classical computers, one could still create simulated humans in the form of biological brains and hook them up to virtual-reality interfaces, like in *The Matrix*. Of course, there might be difficulties with this approach too. For instance, a body laying stationary to receive virtual-reality inputs wouldn't change the brain via exercise in the way that a real biological human's body does. Perhaps the effects of movement and exercise on the brain could be added in without too much difficulty, but maybe not. So there are at least some scenarios in which it would be computationally intractable to simulate a brain in enough detail for it to mirror even just the high-level functional behavior of a biological brain.

A brute-force solution to the above difficulties could be to convert an entire planet to resemble Earth, put real bacteria, fungi, plants, animals, and humans on that planet, and fake signals from outer space (a *Truman Show* approach to simulations), but this would be extremely wasteful of planetary resources (i.e., it would require a whole planet just to run one simulation), so I doubt many advanced civilizations would do it.

Even if simulations can't reproduce the high-level functional behavior of a biological mind, there remains the question of whether some simulations can be made "subjectively indistinguishable" from a biological human brain in the sense that the brain can't tell which kind of algorithm it is, even if the simulation isn't functionally identical to the original biological version. I suspect that this is possible, since the algorithms that we use to reflect on ourselves and our place in the world don't seem beyond the reach of classical computation and indeed may be not insanely complicated. But I suppose it's possible that computationally demanding quantum algorithms are somehow required in this process.

biological humans on Earth and (2) simulated humans who think they're on Earth who are all the same as each other but who are different than biological humans. This is illustrated in the following figure, where the B's represent biological humans and the S's represent the simulated humans who all share the same cognitive algorithm as each other.



Given uncertainty between whether you're aBor anS, you apportion 1/2 odds to being either algorithm. If you're aB, you can influence all N expected sent-years of computation in your future, while if you're anS, you can only influence E sent-years, but there are many copies of you. The calculation ends up being the same as in the "Calculation based on all your copies" section above, since

> L = (probability you're a B) · (number of biological copies of you) · (expected value per copy) + (probability you're an S) · (no impact for future-focused work because there is no far future in a simulation) = (1/2) · (f_y· C · f_C· E · f_E) · (N · T · D) + $\frac{1}{2}$ · 0,

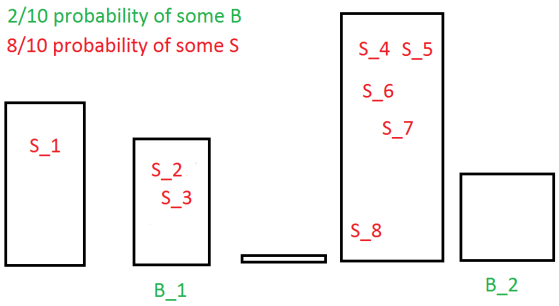
and

> S = (probability you're a B) · (number of biological copies of you) · (expected value per copy) + (probability you're an S) · (number of non-solipsish simulated copies of you) · (expected value per copy) = $\frac{1}{2}$ · (f_y · C · f_C · E · f_E) · E + $\frac{1}{2}$ · (f_y· C · N · f_N) · E

L/S turns out to be exactly the same as before, after we cancel the factors of 1/2 in the numerator and denominator.[13]

Next, suppose that all the simulated copies are different from one another, so that it's no longer the case that what one copy does, the rest do. In this case, there are lots of algorithms that you might be (labelledS_1,S_2, ... in the below figure), and most of them are simulated.

[13]In this setting, it may no longer be reasonable to assume that $f_N = f_X \cdot f_C$, as I did in a previous section, because f_C is the fraction of all civilizations that has the B algorithms on the home planet, while f_N is the fraction of advanced computing power devoted to S algorithms. Since B and S are different algorithms, it may be less plausible that, e.g., if B's are twice as numerous, then S's will be twice as numerous. Nonetheless, since B's and S's are similar enough that you can't tell which you are with your limited reasoning abilities, it may still be somewhat plausible that f_C and f_N are strongly correlated. For instance, even if it's not possible to accurately simulate B algorithms because they involve hard-to-compute quantum effects, it still might be the case that there are S algorithms that are non-quantum-accurate versions of B, and if B algorithms are very common on biological planets, then S algorithms should presumably be very common in simulations.



Now the probability that you're biological is just P_b , and the L/S calculation proceeds identically to what was done in the "Calculation using Bostrom-style anthropics and causal decision theory" section above.

So no matter how we slice things, we seem to get the exact same expression for L/S. I haven't checked that this works in all cases, but the finding seems fairly robust.

8 Solipsist and solipsish simulations

> Since it is harder to vary the simulation detail in role-playing simulations containing real people [i.e., since people are particularly expensive to simulate compared with coarse-grained models of inanimate objects], these simulations tend to have some boundaries in space and time at which the simulation ends. – Robin Hanson

Does consideration of simulations favor solipsist scenarios? In particular, it's possible to run $\sim$ 7billion times more simulations in which you are the only mind than it is to run a simulation containing all of the world's human population. In those superintelligent civilizations where you are run a lot more than average, you have many more copies than normal. So should you be more selfish on this account, since other people (especially distant people whom you don't observe) may not exist?

Maybe slightly. Robin Hanson:

> And your motivation to save for retirement, or to help the poor in Ethiopia, might be muted by realizing that in your simulation you will never retire and there is no Ethiopia.

However, we shouldn't give too much weight to solipsist simulations. Maybe there are some superintelligences that simulate just copies of you. But there may also be superintelligences that simulate just copies of other people and not you. Superintelligences that simulate huge numbers of just you are probably rare. In contrast, superintelligences that simulate a diverse range of people, one of which may be you, are probably a lot more common. So you may have many more non-solipsist copies than solipsist copies.

You may also have many solipsish copies, depending on the relative frequency of solipsish vs. non-solipsish simulations. Solipsish simulations that don't simulate (non-pet) animals in much detail can be much cheaper than those that do, so it's possible there are, say, 5 or 20 times as many solipsish simulations that omit animals than those that contain animals? It's very hard to say exactly, since it depends on the relative usefulness or intrinsic value that various superintelligent simulators place on various degrees of simulation detail and realism. Still, as long as the number of animal-free solipsish simulations isn't many orders of magnitude higher than the number of animal-containing simulations, helping animals is still probably very important.

And the possibility of animal-free solipsish simulations doesn't dramatically upshift the importance of helping developing-world humans relative to helping animals, since in some solipsish simulations, developing-world humans don't exist either.

The possibility of solipsish simulations may be the first ever good justification for giving (slightly) more moral weight to those near to oneself and those one can observe directly.

8.1 Famous people

Jaan Tallinn and Elon Musk both find it likely that they're in a simulation. Ironically, this belief may be more justified for interesting tech millionaires/billionaires than for ordinary people (in the sense that famous/rich people may have more copies than ordinary people do), since it may be both more scientifically useful and more entertaining to simulate powerful people rather than, e.g., African farmers.

So should rich and powerful people be more selfish than average, because they may have more simulated copies than average? Probably not, because powerful people can also make more altruistic impact than average, and at less personal cost to themselves. (Indeed, helping others may make oneself happier in the long run anyway.) It's pretty rare for wealthy humans to experience torture-level suffering (except maybe in some situations at the end of life – in which case, physician-assisted suicide seems like a good idea), so the amount of moral good to be done by focusing on oneself seems small even if most of one's copies are solipsist.

8.2 How feasible are solipsist simulations?

It may be hard to fake personal interactions with other humans without actually simulating those other humans. So probably at least your friends and family are being simulated too. But the behavior of your acquaintances would be more believable if *they* also interacted with fully simulated people. Ultimately, it might be easiest just to simulate the whole world all at once rather than simulating pieces and fudging what happens around the edges. I would guess that most simulations requiring a high level of accuracy contain all human minds who exist at any given time on Earth (though not necessarily at past and future times).

If there are disconnected subgraphs within the world's social network, it's possible there could be a solipsish simulation of just your subgraph, but it's not clear there are many disconnected subgraphs in practice (except for tiny ones, like isolated peoples in the Amazon), and it's not clear why the simulators would choose to only simulate $\sim 99\%$ of the human population instead of 100%.

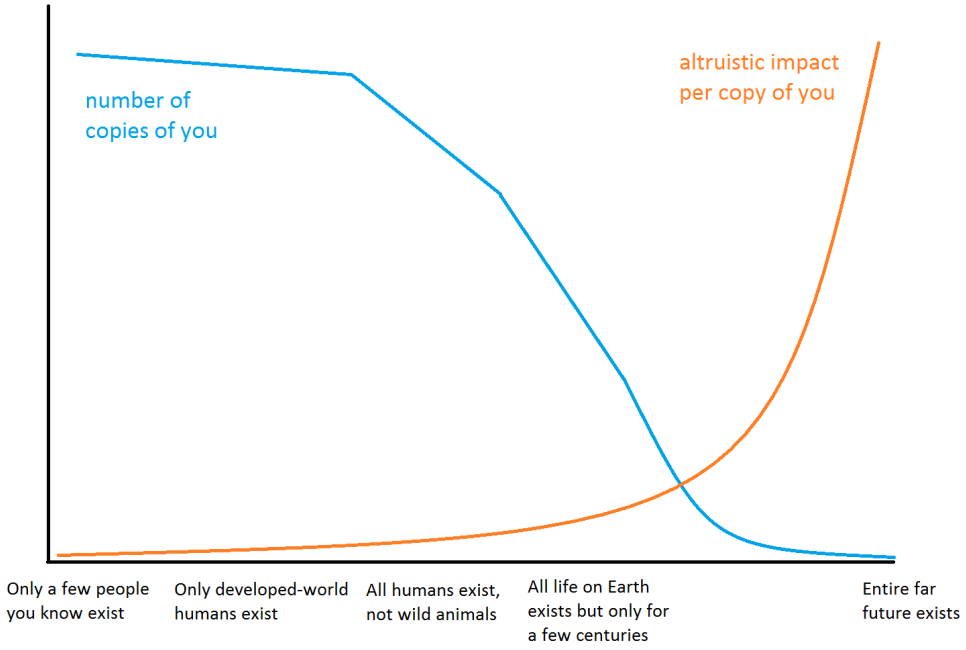
What about non-human animals? At least pets, farm animals, and macroscopic wildlife would probably need to be simulated for purposes of realism, at least when they're being watched. (Maybe this is the first ever good argument against real-time wildlife monitoring and CCTV in factory farms.) And ecosystem dynamics will be more believable and realistic if all animals are simulated. So we have some reason to suspect that wild animals are simulated as well. However, there's some uncertainty about this; for instance, maybe the simulators can get away with pretty crude simulation of large-scale ecosystem processes like phytoplankton growth and underground decomposition. Or maybe they can use cached results from previous simulations. But an accurate simulation might need to simulate every living cell on the planet, as well as some basic physical features of the Earth's crust.

That said, we should in general expect to have more copies in lower-resolution simulations, since it's possible to run more low-res than high-res simulations.

8.3 Tradeoff between number of copies vs. impact per copy

The following figure illustrates some general trends that we might expect to find regarding the number of copies we have of various sorts. Altruistic impact is highest when we focus on the level of solipsishness where the product of the two curves is highest. The main point of this essay is that where that maximum occurs is not obvious. Note that

this graph can make sense even if you give the simulation hypothesis low probability, since you can convert "number of copies of you" into "expected number of copies of you", i.e., (number of copies of you if simulations are common) · (probability simulations are common).



If it turns out that solipsish simulations are pretty inaccurate and so can't reproduce the input-output behavior that your brain has in more realistic worlds, then you won't have copies at all levels of detail along the solipsish spectrum, but you should still have uncertainty about whether your algorithm is instantiated in a more or less long-lived high-resolution simulation, or not in a simulation at all.

9 Suffering in physics or other black swans could save future fanaticism

In this piece, I've been assuming that most of the suffering in the far future that we might reduce would take the form of intelligent computational agents run by superintelligences. The more computing power these superintelligences have, the more sentient minds they'll create, and the more simulations of humans on Earth some of them will also create.

But what if most of the impact of actions targeting the future doesn't come from effects on intelligent computations but rather from something else much more significant? One example could be if we considered suffering in fundamental physics to be extremely morally important in aggregate over the long-term future of our light cone. If there's a way to permanently modify the nature of fundamental physics in a way that wouldn't happen naturally (or at least wouldn't happen naturally for googol-scale lengths of time), it might be possible to change the amount of suffering in physics essentially forever (or at least for googol-scale lengths of time), which might swamp all other changes that one could accomplish. No number of mirrored good deeds across tons of simulations could compete (assuming one cares enough about fundamental physics compared with other things).

Another even more implausible scenario in which far-future focus would be astronomically more important than short-term focus is the following. Suppose that advanced

civilizations discover ways to run insane amounts of computation – so much computation that they can simulate all interesting variations of early biological planets that they could ever want to explore with just a tiny fraction of their computing resources. In this case, F could be extremely small because there may be diminishing returns to additional simulations, and the superintelligences instead devote the rest of their enormous computing resources toward other things. However, one counterargument to this scenario is that a tiny fraction of civilizations might *intrinsically value* running ancestor simulations of their own and/or other civilizations, and in this case, the fraction of all computation devoted to such simulations might not be driven close to zero if obscene amounts of computing power became available. So it seems that F has a lower bound of roughly (computational-power-weighted fraction of civilizations that intrinsically value ancestor simulations) · (fraction of their computing resources spent on such simulations). Intuitively, I would guess that this bound would likely not be smaller than 10^{-15} or 10^{-20} or something. (For instance, probably at least one person out of humanity's current $\sim 10^1$ people would, sadly in my view, intrinsically value accurate ancestor simulations.)

10 The value of further research

This essay has argued that we shouldn't rule out the possibility that short-term-focused actions like reducing wild-animal suffering over the next few decades in terrestrial ecosystems may have astronomical value. However, we can't easily draw conclusions yet, so this essay should not be taken as a blank check to just focus on reducing short-term suffering without further exploration. Indeed, arguments like this wouldn't have been discovered without thinking about the far future.

Until we know more, I personally favor doing a mix of short-term work, far-future work, and meta-level research about questions like this one. However, as this piece suggests, a purely risk-neutral expected-value maximizer might be inclined to favor mostly far-future work, since even in light of the simulation argument, far-future focus tentatively looks to have somewhat higher expected value. The value of information of further research on the decision of whether to focus more on the short term or far future seems quite high.

11 Acknowledgements

Carl Shulman inspired several points in this piece and gave extensive feedback on the final version. My thinking has also benefited from discussions with Jonah Sinick, Nick Beckstead, Tobias Baumann, and others.

References

Armstrong, S. (2011). Anthropic decision theory. *arXiv preprint arXiv:1110.6437*.

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

Bostrom, N. (2003b). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(03):308–314.

Bostrom, N. (2006). Quantity of experience: brain-duplication and degrees of consciousness. *Minds and Machines*, 16(2):185–200.

Bostrom, N. (2010). *Anthropic bias: Observation selection effects in science and philosophy*. Routledge.

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

Lloyd, S. (2006). *Programming the universe: a quantum computer scientist takes on the cosmos*. Vintage Books.