The Washout Argument Against Longtermism

Eric Schwitzgebel
Department of Philosophy
University of California, Riverside
Riverside, CA 92521
USA

December 27, 2023

The Washout Argument Against Longtermism

Abstract: We cannot be justified in believing that any actions currently available to us will have a non-negligible positive influence on the billion-plus-year future. I offer three arguments for this thesis. According to the Infinite Washout Argument, standard decision-theoretic calculation schemes fail if there is no temporal discounting of the consequences we are willing to consider. Given the non-zero chance that the effects of your actions will produce infinitely many unpredictable bad and good effects, any finite effects will be washed out in expectation by those infinitudes. According to the Cluelessness Argument, we cannot justifiably guess what actions, among those currently available to us, are relatively more or less likely to have positive effects after a billion years. We cannot be justified, for example, in thinking that nuclear war or human extinction would be more likely to have bad than good consequences in a billion years. According to the Negligibility Argument, even if we could justifiably guess that some particular action is likelier to have good than bad consequences in a billion years, the odds of good consequences would be negligibly tiny due to the compounding of probabilities over time.

Word count: ~6500 words

Keywords: decision theory; futurism; longtermism; MacAskill, William

The Washout Argument Against Longtermism

According to *longtermism*, our decisions should be substantially influenced by our expectations for their effects on the long-term future – not just the next ten years or even the next thousand years, but the next billion-plus years. MacAskill's 2022 book *What We Owe the Future* is perhaps the best-developed articulation of this view (see also Beckstead 2019; Ord 2020; Greaves and MacAskill 2021; Tarsney 2023). I will argue, on the contrary, that our decisions should be *not at all* influenced by our expectations about their effects more than a billion years in the future.

My grounds are this: There are no practically available actions – nothing we can actually do now – that we are justified in believing will have a non-negligible positive influence on events more than a billion years from now, due to the many massively complex possible causal ramifications of any action. Partly following Greaves and MacAskill (2021), I will call this the *washout argument* against longtermism.[1] I offer three versions of the washout argument:

- an infinite version, which explores the unintuitive consequences of granting even a small chance that our actions have infinitely many consequences;
- a finite version, which asserts cluelessness about what current actions would have good versus bad impacts after a billion years;
- another finite version, which asserts that even if we could justifiably guess what current actions would have good versus bad impacts in a billion-plus years, the expected impact of any currently available action is negligible.

[1] This definition of washout is somewhat broader than "washing out" as it is sometimes used in the literature, encompassing Greaves and MacAskill's (2021) "washing out" and "cluelessness", Thorstad's (2021) "rapid diminution" and "washing out", and Tarsney's (2023) "epistemic challenge".

I will conclude by cheering longtermists' practical suggestions for devoting resources to the *medium-term* (several-hundred-year) future.

1. Infinite Washout.

The most straightforward interpretation of current physical theory appears to imply that the universe will endure infinitely and that almost all of our actions will have infinitely many positive and negative effects into the future. I have defended this view in detail elsewhere (Schwitzgebel and Barandes forthcoming). I will only sketch the general idea here. If you raise your hand now, that will cause many particles to change their trajectories, which will then cause many other particles to change their trajectories, which will then cause many other particles to change their trajectories, in an ever-expanding ripple of effects that extends into the post-heat-death universe. Since there's no reason to expect a future temporal boundary at which time ceases, a reasonable default assumption is that the universe will endure infinitely post-heat-death. Most physical theories imply there will occasionally be chance thermodynamic or quantum fluctuations of very large magnitude, and there appears to be no cap on the size of these fluctuations. If so, then given infinite time to wait, whole new galaxies will eventually fluctuate into existence by massively improbable chance. Alternatively, black holes or some other process might seed new cosmic inflations, giving rise to new galaxies by those more ordinary-seeming means. Either way, the still-rippling effects of your current actions will then influence what occurs in those galaxies. Eventually – maybe after a googolplex-to-the-googolplex years – a photon that wouldn't have been exactly where it is, with exactly the energy it in fact has if I hadn't raised my hand just now, will strike a future radioastronomer's device in just the right way to push a triggering mechanism over a threshold, causing the astronomer to check a device

and then write an article that wins a prize, forever changing her career. Another photon will cause another ripple of effects that gives some future diplomat cancer, resulting in a power shift that precipitates a terrible war. Continuing into the infinite future, if this cosmological model is correct, virtually every action you perform will have infinitely many positive and infinitely many negative effects. Call this the *Infinitary Cosmology.*

This is a problem if your decision procedure requires weighing up all the expected positive and negative effects of your actions, with no temporal discounting. Suppose Action A is giving a million dollars to an effective charity, with benefit m, and Action B is using that money to burn down the house of a neighbor with an annoying dog, with benefit n (n might be negative). The total expected value of Action A might be modeled as $m + \infty + -\infty$. The total expected value of Action B might be modeled as $n + \infty + -\infty$. Neither value is mathematically defined. They can't be weighed against each other. Nor can we escape the problem by assuming, for example, that over the long term, the positive effects outweigh the negative effects. Then the expected value of A would be $m + \infty$ and the expected value of B would be $n + \infty$. These values are equal. As the saying goes, infinity plus one is just infinity.

We also can't escape the problem by assuming that in the long run the positive and negative effects of both actions approximately balance. As long as positive and negative effects are randomly distributed, in both the Action A stream and the Action B stream, there will be no finite limit to the size of a positive or negative run of good or bad effects in either stream: Neither series will converge to a zero sum. What might converge to zero is the *ratio* of positive to negative effects as the series of effects goes to infinity. But then again, if we're looking at ratios rather than sums, there's nothing to choose between Actions A and B: Both converge

toward an equal ratio of positive vs. negative effects: In the ratio limit, the initial *m* and *n* wash out.

Now you might regard the Infinitary Cosmology as unlikely. Maybe you even have a high credence that it's false – say, 99.9% credence. The possible infinitudes will still destroy your long-term expected value calculations. Suppose that you know that if the Infinitary Cosmology is false, the expected value of A will be *m*. Given your 0.1% credence in the Infinitary Cosmology, the expected value of your action will be .999 * *m* + .001 * (∞ + -∞) – in total, an undefined value.

You might think that symmetry or dominance reasoning can help you escape this conclusion. Not so. Consider the following. Eventually, among the long-term effects of your million-dollar charitable donation (value *m*) will be the value equivalent of burning down your neighbor's house (value *n*). Eventually, among the long-term effects of your burning down the house (value *n*) will be the value equivalent of donating a million dollars to charity (value *m*). Of course, there will still remain infinitely many positive and negative effects of both actions: Pulling an *n* or an *m* out of the infinitude will not make it any less an infinitude. Thus, we can model the effects of A and B as equally *m* + *n* + ∞ + -∞. Action A and B can thus be modeled as symmetric, with neither dominating the other. In general, trying to subtract away infinitudes, or balance them against each other, or engage in other linear transformations of infinitude, leads to paradox unless we confine ourselves carefully to the mathematics of limits. Maybe this is unintuitive. But infinitude is unintuitive.

If there's a non-zero chance that your actions have infinitely many positive and negative effects in an Infinitary Cosmology of the sort described, then unless there's some temporal discounting, standard decision theory collapses. The lesson I draw is that there's a problem with

one of the background assumptions of longtermism as formulated by MacAskill and others. Longtermists' application of standard decision theory with no temporal discounting will inevitably generate either indifference, paradox, or undefined values. The tools break. We must think about our decisions in some other way.[2]

2. Finite Washout: Cluelessness.

Let's suppose that the longtermist ducks the Infinite Washout objection by implementing a temporal cutoff – say, at the heat death of the universe. They might justify this cutoff by pleading radical ignorance about good or bad effects after heat death: The good or bad effects of our current actions on any particular spatiotemporal region post heat death will be sheer unpredictable chance, not worth figuring into our decision calculus.

That is a reasonable response. But notice that it is already a significant concession. It abandons the time-indifferent calculus that longtermists tend to prefer in favor of discounting the far future. And it requires what is arguably an inelegant and unjustified cutoff at heat death – inelegant and unjustified in part because arguably we should assign some tiny credence to the possibility that our descendants persist even after heat death, for example, by discovering new sources of energy in violation of current understandings of the laws of thermodynamics. Longtermists could perhaps find a more elegant solution by employing a smoother discount

[2] Schwitzgebel (forthcoming-a) also presents a version of this argument; see also Nelson (1991) for a precedent. Bostrom (2011) argues that no formal approach to decision making in an infinite context is likely to be entirely elegant and intuitive, since the mathematics of infinitude can't handle all the plausible cases, and various discounting regimes appear to generate unintuitive consequences. Easwaran (2021) offers a dominance-reasoning based solution for a limited range of cases, but the Infinitary Cosmology does not belong to that range. Wilkinson (2021) also offers no good decision-theoretic solution to the problem, finding the outcomes in cases like charity-vs-house-burning to diverge to *either* $+\infty$ or $-\infty$ with equal probability, and thus no decision-theoretical basis to choose one over the other.

function – but smooth discounting functions generally have other unintuitive consequences and are directly contrary to the spirit in which longtermism is typically offered.[3] Furthermore, if the longtermist appeals to post-heat-death ignorance to justify disregarding consequences that far in the future, they open themselves up to the possibility of an earlier ignorance-based cutoff date.

The last point is the one I will press in this section. We are radically ignorant about the value of our current actions for the state of things a billion years in the future. A billion years is on the neartermish end of the time between now and heat death: Sun won't swallow Earth for about another 8 billion years, and the last stars of the observable portion of the universe aren't estimated to burn out for many trillions of years.

If MacAskill's and most other longtermists' reasoning is correct, the world is likely to be better off in a billion years if human beings don't go extinct now than if human beings do go extinct now, and decisions we make now can have a non-negligible influence on whether that is the case. In the words of Toby Ord, humanity stands at a precipice. If we reduce existential risk now, we set the stage for possibly billions of years of thriving civilization; if we don't, we risk the extinction of intelligent life on Earth. It's a tempting, almost romantic vision of our importance. I also feel drawn to it. But the argument is a card-tower of hand-waving plausibilities. Equally breezy towers can be constructed in favor of human self-extermination or near-self-extermination. Let me offer two.

The Dolphin Argument. The most obvious solution to the Fermi Paradox is also the most depressing. The reason we see no signs of intelligent life elsewhere in the universe is that technological civilizations tend to self-destruct in short order. If technological civilizations tend to gain increasing destructive power over time, and if their habitable environments can be

[3] See discussions in Bostrom 2011 and Ord 2020.

rendered uninhabitable by a single catastrophic miscalculation or a single suicidal impulse by someone with their finger on the button, then the odds of self-destruction will be non-trivial, might continue to escalate over time, and might cumulatively approach nearly 100% over millennia. I don't want to commit to the truth of such a pessimistic view, but in comparison, other solutions seem like wishful thinking – for example, that the evolution of intelligence requires stupendously special circumstances (the Rare Earth Hypothesis) or that technological civilizations are out there but sheltering us from knowledge of them until we're sufficiently mature (the Zoo Hypothesis).

Anyone who has had the good fortune to see dolphins at play will probably agree with me that dolphins are capable of experiencing substantial pleasure. They have lives worth living, and their death is a loss. It would be a shame if we drove them to extinction.

Suppose it's almost inevitable that we wipe ourselves out in the next 10,000 years. If we extinguish ourselves peacefully now – for example, by ceasing reproduction as recommended by antinatalists – then we leave the planet in decent shape for other species, including dolphins, which might continue to thrive. If we extinguish ourselves through some self-destructive catastrophe – for example, by blanketing the world in nuclear radiation or creating destructive nanotech that converts carbon life into gray goo – then we probably destroy many other species too and maybe render the planet less fit for other complex life. To put some toy numbers on it, in the spirit of longtermist calculation, suppose that a planet with humans and other thriving species is worth X utility per year, a planet with other thriving species with no humans is worth X/100 utility (generously assuming that humans contribute 99% of the value to the planet!), and a planet damaged by a catastrophic human self-destructive event is worth an expected X/200 utility. If we destroy ourselves in 10,000 years, the billion year sum of utility is $10^4 * X +$

(approx.) 10^9 * X/200 = (approx.) 5 * 10^6 * X. If we peacefully bow out now, the sum is 10^9 * X/100 = 10^7 * X. Given these toy numbers and a billion-year, non-human-centric perspective, the best thing would be humanity's peaceful exit.

Now the longtermists will emphasize that there's a chance we won't wipe ourselves out in a terribly destructive catastrophe in the next 10,000 years; and even if it's only a small chance, the benefits could be so huge that it's worth risking the dolphins. But this reasoning ignores a counterbalancing chance: That if human beings stepped out of the way a *better* species might evolve on Earth. Cosmological evidence suggests that technological civilizations are rare; but it doesn't follow that *civilizations* are rare. There has been a general tendency on Earth, over long, evolutionary time scales, for the emergence of species with moderately high intelligence. This tendency toward increasing intelligence might continue. We might imagine the emergence of a highly intelligent, creative species that is less destructively Promethean than we are – one that values play, art, games, and love rather more than we do, and technology, conquering, and destruction rather less – descendants of dolphins or bonobos, perhaps. Such a species might have lives every bit as good as ours (less visible to any ephemeral high-tech civilizations that might be watching from distant stars), and they and any like-minded descendants might have a better chance of surviving for a billion years than species like ours who toy with self-destructive power. The best chance for Earth to host such a species might, then, be for us humans to step out of the way as expeditiously as possible, before we do too much harm to complex species that are already partway down this path. Think of it this way: Which is the likelier path to a billion-year happy, intelligent species: that we self-destructive humans manage to keep our fingers off the button century after century after century somehow for ten million centuries, or that some other

more peaceable, less technological clade finds a non-destructive stable equilibrium? I suspect we flatter ourselves if we think it's the former.

This argument generalizes to other planets that our descendants might colonize in other star systems. If there's even a 0.01% chance per century that our descendants in Star System X happen to destroy themselves in a way that ruins valuable and much more durable forms of life already growing in Star System X, then it would be best overall for them never to have meddled, and best for us to fade peacefully into extinction rather than risk producing descendants who will expose other star systems to their destructive touch.

The Nuclear Catastrophe Argument. MacAskill and other longtermists emphasize the importance of humans learning to take existential risks seriously. They also generally hold that a nuclear war would not likely permanently destroy the species. This suggests a method for teaching humanity to take existential risk seriously: Start a nuclear war.

Suppose that a nuclear war has a 5% chance of destroying humanity, but that if it doesn't destroy us, for the next 10,000 years the survivors take existential risk more seriously, reducing the existential risk to humanity from, say 2% per century to 1.9% per century. Assuming that risk per century is otherwise constant and independent, the 10,000-year survival odds are as follows: without nuclear war, 13.3%; with nuclear war, 14.0%. Now might in fact be the optimal time to engage in all-out nuclear war. If our weapons of mass destruction grow more powerful over the next hundred years, all-out war will likely be riskier. Our odds of surviving all-out worldwide war in a hundred years might only be 75% instead of 95%. Best, then, to act in the sweet spot where there's enough destruction to durably scare us into a sober attitude, while it's still very likely we'd survive!

Is this implausible? I don't think so. Humans learn much better by hard experience than by philosophical treatises about relative risks and benefits. Humanity might be like the reckless teen driver whose nearly fatal accident finally teaches them caution. Humanity is far too cavalier about existential risk, longtermists say. Well, what could teach us more powerfully and unforgettably?

My aim with the Dolphin Argument and the Nuclear Catastrophe Argument is not to convince readers that humanity should bow out for the sake of other species, much less that we should start a nuclear war. Rather, my thought is this: It's easy to concoct stories about how what we do now might affect the billion-year future, and then to attach decision-theoretic numbers to those stories. We lack good means for evaluating these stories. We are likely just drawn to one story or another based on what it pleases us to think and what ignites our imagination.

I suggest that it is *no more plausible* that the best thing we can do for the billion-year future is, as MacAskill suggests, fund studies of existential risk in hopes that we can permanently lock in positive values than that the best thing we can do for the state of the world in a billion years is suffer near-term extinction or global catastrophe.

I'm not saying that extinction or nuclear war (or whatever) are *almost* as likely to have good as bad billion-year consequences. If our credences are close to parity but uneven – say, 50.1% good vs. 49.9% bad – the longtermist argument can get a toehold by appeal to extremely high stakes (50.1% times trillions is much better than 49.9% times trillions). I am making a stronger claim of radical uncertainty. One way of characterizing this radical uncertainty is to balance the credences and values precisely. Maybe for all values of x, contingent on assuming that the world is either better or worse by x amount over the course of a billion years as a result

of having a nuclear war now, our credence that the world is better by x amount should be exactly 50%. Or maybe better: If the expected short-to-medium term impact of an action is y, we should always treat the expected billion-year-impact as also y. Alternatively, maybe the best approach to radical uncertainty is to say that here, as in the infinite case, mathematical decision theory fails – more likely to mislead than enlighten.

We are radically clueless about what we can do now that would be likely to have good effects after a billion years. We can reasonably estimate *some* things about the billon-year future. It's a good guess that the Sun will still exist. But the positive or negative utilities in Year One Billion of current efforts to prevent human extinction or near-extinction wash out in a fog of radical inscrutability.[4]

3. Finite Washout: Negligibility.

Suppose we ignore the problem of infinite utilities, and suppose we also reassuringly guess that various things we can do now can reasonably be estimated to improve the long-term future by improving the odds of continued human existence. The question arises: *How much* impact on the far-distant future should we expect from any action that is possible today? There's reason to think that the expected impact would be negligibly small.

To get a sense of how small, let's start with some toy numbers. Suppose that I can do some Action A now (perhaps a $100,000 donation to the Longtermism Fund) with the following

[4] Although the appeal to cluelessness about long-term effects might bring to mind Lenman (2000), I accept Greaves's (2016) critique of the Lenman argument for finite cases. Perhaps the dolphin and nuclear war cases can be treated as the more intractable form of "complex cluelessness" described by Greaves; see also Mogensen (2021) on the complexity of comparing donation to Make-A-Wish vs. Against Malaria. The Infinite Washout argument, however, can be seen as an adaptation of Lenman's argument for a particular infinite cosmological possibility.

expected consequences in the temporal range from Year One Billion to heat death: a one in a trillion chance of leading to a trillion happy descendants who wouldn't otherwise exist, plus a one in ten trillion chance of ten trillion happy descendants, plus a one in a hundred trillion chance of a hundred trillion happy descendants, and so on. Assuming for simplicity that all descendants' lives are of equal value and applying the expected formula in the usual way, Action A would have infinite value. Ah – but wait, that would presumably require infinite resources, and this is supposed to be the finite portion of the argument. So let's cap the possible number of happy descendants. If we cap at 10^{16}, and if we value one future happy life that wouldn't otherwise exist at one util, then the expected value of Action A is 1 + 1 + 1 + 1 + 1 = 5 utils. If we more generously cap the possible number of happy descendants at 10^{50}, then the value is 39 utils – quite possibly worth my $100,000 if there's no temporal discounting.

But tweaking the probabilities yields very different numbers. Suppose that Action A has a one in ten quadrillion chance of leading to a trillion happy descendants who wouldn't otherwise exist, a one in 200 quadrillion chance of leading to ten trillion happy descendants who wouldn't otherwise exist, a one in four quintillion chance of leading to a hundred trillion happy descendants, and so on. Now, even with no cap on the number of descendants, the expected value of Action A is 1/10000 + 1/20000 + 1/40000 + 1/80000 + … = 2/10000 descendants – less than one thousandth of a happy life. You might do much better spending that $100,000 on mosquito nets, or maybe just having some unprotected sex.

So, which set of numbers is more reasonable? For comparison, consider the chance of changing the outcome of an election by voting. Political scientists have estimated this in various ways, but in a simple popular-vote election with ten million voters, a not-unreasonable ballpark estimate is that there's a one in ten million chance that your vote is decisive (Gelman, Katz, and

Bafumi 2004; Barnett 2020). I suggest that the confidence you should have that any action you perform now will lead to a trillion or more additional good lives in the long-term future should be *many orders of magnitude less* than the confidence you should have that your vote in a ten-million person election will be the decisive vote for your favored candidate. A one in a trillion chance – what we assumed in our first calculation – seems far too close in magnitude to one in ten million, only five orders of magnitude different, despite an intervening billion years. In a billion years, a lot can happen to ruin your plans. Before any action of yours could give rise to a trillion descendants, there must be at least 999 billion other people with various other plans. All of these many billions of people will be temporally closer to that billion-year future than you are and thus presumably able to influence that future more directly than you can, with fewer intervening causal threads and fewer contingencies that need to work out favorably.

You cast your stone into the pond, hoping for a billion-year positive ripple. You donate that $100,000. Suppose you estimate a 50% chance that in a hundred years your action is still having some positive consequences, versus a 50% chance that it is washed out, counteracted, or lost into the noise of the past, with no continuing overall positive influence of the sort you were hoping for. (A washed out action, in this sense, might still have random or unpredictable positive and negative ripples in the manner described in Section 1, but it would take a great stroke of luck for all those ripples to add up to plus-one-trillion good lives, and once the effects become random, symmetry suggests similar odds of similar magnitude negative effects.) Will it continue to have the positive effects you were hoping for in another hundred years, and another, and another? If we assume a 50% washout chance per hundred years, the odds that your action is still having those overall positive consequences in a billion years is one in $2^{10,000,000}$ – that is, under one in $10^{1,000,000}$.

Presumably with calculations of this sort in mind, MacAskill, Ord, and other longtermists argue that we are currently in a *uniquely crucial* time. If by surviving the next 10,000 years we could almost eliminate existential risk and maybe also permanently lock in good values, then our actions now are particularly significant and stand a reasonable chance of not being washed out in the ten million centuries between now and a billion years from now.

It's not completely implausible to think that we live in a uniquely crucial time. Maybe existential risk would fall dramatically if we were to colonize other star systems sufficiently independent from one another that a catastrophe in one is unlikely to precipitate catastrophe throughout. Maybe we're now in a uniquely risky ten-thousand-year window, pre-colonization. Or maybe we have a unique opportunity over the next few millennia to durably lock in a set of future values by using emerging AI to create a rigid, unchangeable, billion-year-long world governance system based on our current values. To me, this seems somewhat less likely than space colonization. If the system is too inflexible, that might increase existential risk via inability to respond flexibly to unforeseen problems. More fundamentally, if future generations have the kinds of creative, intelligent capacities that human philosophers tend to prize and that presumably the longtermists want to preserve, it seems unlikely that we can bind them forever to our current value systems.

But suppose we grant a 1% chance of the near-elimination of existential risk and the enduring lock-in of good values if we survive the next 10,000 years, after which there's just a 1% chance per million years of extinction or some other type of washout of your current good intentions. A 99% chance of continuing good effects per million years for a thousand million-year eras compounds to $(0.99)^{1000}$ – only a 0.004% chance of continuing good effects a billion years out. Even under these generous assumptions, that's still a seven-orders-of-magnitude

decrement, atop whatever small chance there is of your action continuing its intended good effects over the first 10,000 years. It's hard to see how any action you take now could realistically have anything like a one in a trillion chance of leading to a trillion more good lives in a billion years.

To think otherwise is, I suspect, to not vividly appreciate how long a billion years really is or how quickly chances multiply. It only takes twelve independent one-in-ten chances to fall to odds of one in a trillion. It seems likely that for any action of yours to cause a trillion more good lives to exist a billion years or more in the future would require much more than twelve independent one-in-ten chances to land your way. Suppose, instead, it requires a thousand one-in-ten chances. Maybe every million years, something a little lucky needs to happen. To me, even this estimate seems extremely optimistic. The expected value of that action is then $10^{12}/10^{1000}$ – well under a googolth of a life. Adding further one-in-ten chances of bigger impacts – say, another 1 in 10 chance for every extra order of magnitude of lives up to a maximum of 10^{50} ($10^{13}/10^{1001}$, $10^{14}/10^{1002}$, etc.) – still doesn't bring the total expectation above one in a googolth. Only if we calculate long past heat death will we get a different result. But unless there's some principled reason to start discounting long after heat death, such a move risks casting us back into the problems with the Infinitary Cosmology.[5]

[5] One might think of the longtermist argument as a Pascal's mugging argument in disguise (Bostrom 2009). Pascal's mugging only succeeds if your credence in the mugger's promised outcomes does not decrease sufficiently quickly to offset the additional goods promised. For example, if your credence that the mugger will later give you $1 is 1/1000, that he will give you $2 is 1/4000, that he will give you $3 is 1/8000, etc., you won't pay him even a penny. A reasonable response to both Pascal's mugger and longtermist speculations about 10^{30} or more happy future lives is to decrease one's credence sharply with each additional order of magnitude of hypothesized benefit. (Another solution is to generally reject standard decision-making models under conditions of extremely low probabilities of extremely high or low value outcomes, as discussed in Monton 2019 and applied to longtermism in Tarsney 2023.)

I'm not committed to particular odds or magnitudes, other than the general observation that the odds are tiny. My main thought is, I hope, sufficiently plausible that we should accept it absent some rigorous argument to the contrary. Any action you engage in now has only a minuscule chance of making the world vastly better in a billion years, and the chance grows precipitously more minuscule the larger the imagined benefit. Our decision-making should be grounded in nearer-term consequences. Whatever you draw in the beach sand today has no material chance of enduring a billion years; and even if the sands tumble around differently evermore, due to chaotic ripple effects upon ripple effects, we cannot possibly know whether these effects will sum more toward the good or the bad.[6]

The Infinite Washout argument establishes the incoherence of standard longtermist decision theory without relying on any particular empirical facts, as long as we have a non-zero credence in the Infinitary Cosmology. In contrast, the finite versions of the washout argument do depend on particular empirical facts. For example, if there were some action available to us now that would entirely obliterate planet Earth, then I would concede that it's reasonable to think there's a good chance that the billion-year future would be much worse if we perform that action. (One might doubt even this, though, if one worries that future life might be so miserable, on average, that it's better to just end things now.) Thus, unlike Infinite Washout, Cluelessness and Negligibility are not arguments against longtermism in principle, regardless of situation or background knowledge. They are only arguments that *we ourselves* should not be longtermists.

4. The Costs of Longtermist Thinking.

[6] See Thorstad (2021, 2023a,b) for more detailed critiques of some of the numerical assumptions underlying longtermist arguments.

The argument of Section 3, if successful, only established that billion-plus-year expectations should have at most a very small influence on our decisions. But suppose we rationally come to believe that Options A and B are identical in their less-than-billion-year expected value while Option A has a better billion-year-plus expected value? Could we then rationally use those very long term expectations to justify selecting Option A?

A case can be made against even that seemingly innocuous application of longtermist reasoning if there's reason to think that longtermist reasoning is costly. I see three ways in which longtermist reasoning might be costly:

First, longtermist reasoning might be effortful. If the expected benefit of longtermist reasoning is a quadrillionth of a life, it might not be worth even a millisecond of cognitive effort to try to get the decision right; one's cognitive resources are better expended elsewhere.[7]

Second, longtermist reasoning might have negative effects on other aspects of one's cognitive life, for example, by encouraging inegalitarian or authoritarian fantasies, or a harmful neo-liberal quantification of goods, or self-indulgent rationalization, or a style of consequentialist thinking that undervalues social relationships or already suffering people (Srinivasan 2015; Torres 2021).

Third, given its difficulty and uncertainty, billion-year-plus thinking adds significant noise and risk of error to our decision making. For example, overconfidence is endemic in forecasting, and even expert geopolitical and economic forecasters tend to decline toward chance over five- to ten-year time frames (Tetlock and Gardner 2015). It's a reasonable conjecture that decisions driven by a billion-year-plus forecasts, to the extent they diverge from decisions driven

[7] See Schwitzgebel forthcoming-b, ch. 4, for a similar argument in favor of disregarding outcomes with less than a 10^{30} chance of occurring, if one has not already paid the cognitive cost of considering them.

by nearer-term forecasts, will tend to be mistakes of overconfidence. On higher-order epistemic grounds, we should avoid billion-year longtermist reasoning as likelier to lead to costly mistakes than to generate valuable insight.

The first and third claims seem plausible on their face, and the second seems plausible if the third is plausible, since cognitive complexity is fertile soil for cognitive vice. If any of these three claims about the costs of longtermist thinking is correct, they can be appended to earlier considerations about negligibility to derive the conclusion that we shouldn't give any weight to billion-year-plus outcomes in our decision making, even just to resolve apparent medium-term ties. The cost of billion-year thinking outweighs the tiny potential expected benefits.

5. The Medium Term.

A thousand years is more visible than a billion. No evolutionary descendants of dolphins or bonobos will be constructing civilizations within a thousand years, at least not without the help of humans or our descendants or constructions. A nuclear war would plausibly be overall bad in its thousand-year implications. And some people or organizations who exist today might perform actions whose consequences retain a stable overall positive or negative balance a thousand years from now – though if we imagine a medieval scholar attempting to predict the various good and bad thousand-year consequences of agricultural innovations, or of the reunification of China by the Song Dynasty, or of the policies of the medieval Catholic Church, we begin perhaps to see the challenges we face in anticipating the various good and bad thousand-year consequences of, say, developments in Artificial Intelligence or world governance. If the arguments of Sections 2 through 4 are correct, at some point after a thousand years but before a billion years – ten thousand years? a million years? ten million years? – any

available action you could now perform is so swamped over with unpredictable good and bad possible effects that you should give that distant future no weight in your decisions.

Still, a thousand years is long term compared to quarterly profits and next year's elections, and we should probably attempt some humble, uncertain conjectures concerning what would be relatively good or bad for the world over that time frame. Since the specific actions that longtermists like MacAskill and Ord recommend are plausibly good over a thousand-year time frame, I think we can support such actions under that conception. It's worth taking AI risk and climate change seriously, not just for their possible effects on our grandchildren but also for their possible effects in six hundred years. However, let's not try to bolster the case for action by conjecturing about one-in-a-trillion chances of particular good or bad consequences a billion years in the future.[8]

[8] For helpful discussion, thanks to William MacAskill, Brice Bantegnie, Jonathan Birch, Richard Yetter Chappell, Kenny Easwaran, Myisha Cherry, Matthew Southey, David Thorstad, and the various people who have commented on relevant posts on my blog and social media accounts.

References

Barnett, Zach (2020). Why you should vote to change the outcome. *Philosophy & Public Affairs, 48,* 422-446.

Beckstead, Nick (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves and T. Pummer, eds., *Effective altruism.* Oxford University Press.

Bostrom, Nick (2009). Pascal's mugging. *Analysis, 69,* 443-445.

Bostrom, Nick (2011). Infinite ethics. *Analysis and Metaphysics, 10,* 9-59.

Easwaran, Kenny (2021). A new method for value aggregation. *Proceedings of the Aristotelian Society, 121,* 299-326.

Gelman, Andrew, Jonathan N. Katz, and Joseph Bafumi (2004). Standard voting power indexes do not work: An empirical analysis. *British Journal of Political Science, 34,* 657-674.

Greaves, Hilary (2016). XIV – Cluelessness. *Proceedings of the Aristotelian Society, 116,* 311-339.

Greaves, Hilary, and William MacAskill (2021). *The case for strong longtermism.* Global Priorities Institute, working paper no. 5-2021.

Lenman, James (2000). Consequentialism and cluelessness. *Philosophy & Public Affairs, 29,* 342-370.

MacAskill, William (2022). *What we owe the future.* Basic Books.

Monton, Bradley (2019). How to avoid maximizing expected utility. *Philosophers' Imprint, 19* (18).

Mogensen, Andreas L. (2021). Maximal cluelessness. *Philosophical Quarterly, 71,* 141-162.

Nelson, Mark T. (1991). Utilitarian eschatology. *American Philosophical Quarterly, 28,* 339-347.

Ord, Toby (2020). *The precipice.* Hachette.

Schwitzgebel, Eric, and Jacob Barandes (forthcoming). Almost everything you do causes almost everything (under certain not wholly implausible assumptions); or infinite puppetry. In E. Schwitzgebel, *The weirdness of the world.* Princeton University Press.

Schwitzgebel, Eric (forthcoming-a). Repetition and value in an infinite universe. In S. Hetherington, ed., *Extreme Philosophy.* Routledge.

Schwitzgebel, Eric (forthcoming-b). *The weirdness of the world.* Princeton University Press.

Srinivasan, Amia (2015). Stop the robot apocalypse. *London Review of Books, 37* (18), September 24.

Tarsney, Christian (2023). The epistemic challenge to longtermism. *Synthese, 201* (195).

Thorstad, David (2021). *The scope of longtermism.* Global Priorities Institute, working paper no. 6-2021.

Thorstad, David (2023a). High risk, low reward: A challenge to the astronomical value of existential risk mitigation. *Philosophy and Public Affairs, 51,* 373-412.

Thorstad, David (2023b). *Three mistakes in the moral mathematics of existential risk.* Global Priorities Institute, working paper no. 7-2023.

Torres, Émile P. (2021). The dangerous ideas of "longtermism" and "existential risk". *Current Affairs,* July 8.

Wilkinson, Hayden (2021). Infinite aggregation: Expanded addition. *Philosophical Studies, 178,* 1917-1949.