The Hinge of History and the Choice between Patient and Urgent Longtermism

Olle Häggström[1]

Version of December 19, 2022

Abstract. The notion that we are living at the *Hinge of History* is meant to capture the idea that our time is uniquely pivotal in human history. Here, ways to make this precise are discussed, and the idea is defended against William MacAskill's so-called base rate argument for out *not* living at the Hinge of History. Finally, implications for the choice between patient and urgent longtermism – i.e., whether a philanthropist wishing to make the long-term future go well should fund object-level action today or save and invest for future such action – are discussed.

Key words: Hinge of History, patient longtermism, urgent longtermism, existential risk.

1. Introduction

Over the past decade it has become increasingly common among scholars working on existential risk and the long-term future of humanity to point to the present time as uniquely and momentously important for the rest of human history. Holden Karnofsky (2021b) speaks of ours as "the most important century" and "the one that will initiate, and have the opportunity to shape, a future galaxy-wide civilization". Here, in a similar spirit, is Max Tegmark:

> It's now, for the first time in the 4.5 billion years history of this planet, that we are at this fork in the road. It's probably going to be within our lifetimes that we're either going to self-destruct or get our act together. (Harris, Goldstein and Tegmark, 2018, 1:16:40 into the recording.)

The name, coined by Derek Parfit (2011), that has caught on for this idea is that "we live during the Hinge of History". But is that statement true? This was largely taken for granted until the important recent work by William MacAskill (2022a) who addresses this question.[2] After giving the Hinge of History (HoH) concept a (relatively) precise definition, MacAskill offers a probabilistic argument for a 'no' answer: we probably do not live at the HoH, which instead is likely to lie in the far future.

This finding, if it holds up to scrutiny, may have implications for so-called *longtermism*, which we take here to be the view that effects on the very long-term future matters greatly for what we ought to do today. If the HoH is not now but in the future, the best way to promote a good long-term future might not be to take concrete action now, but rather to save and invest our resources until the future true HoH, when we can expect more bang for the buck. Distinguishing between *urgent longtermism*, which recommends taking object-level action now, and *patient longtermism*, which recommends saving and investing now in order to take action later, MacAskill's argument thus speaks in favor of the patient approach.

In the present paper, I offer a partly opposing viewpoint, defending the idea that it is reasonably likely that we live during the HoH, and also arguing against patient longtermism and in favor of a more urgent approach. The paper is organized as follows. In Section 2 I discuss how to define the

[1] olleh@chalmers.se

[2] MacAskill's paper has a precursor in a tentative blog post (MacAskill, 2019a), which however I will not be referring except when required by context, focusing instead on his 2022 paper.

hinginess quantity that is maximized at the true HoH, after which in Section 3 I outline MacAskill's so-called base rate argument for why the present time is probably not the HoH. Then, in Section 4, I go on to show that his argument is not as strong as it may seem, and I argue that on the contrary we have good (albeit far from conclusive) reasons to think that we are in fact living at the HoH. In Section 5 I treat the issue of patient vs urgent longtermism, followed in Section 6 by some concluding remarks.[3]

2. Defining the Hinge of History

The issue addressed in this section boils down to defining some function H(t) – the hinginess at time t – that captures how pivotal time t is for how well the rest of human history will go. We can then define the HoH as the time t at which H(t) attains its largest value throughout human history, and go on to address the problem of whether the HoH is now. In principle, "now" might mean "August 31, 2022" or even be pinned down to a particular time that day, but in the interest of decision relevance or just having a realistic hope of arriving at an interesting answer, it is better to take it to mean something like "the present decade" or "the present century". In connection with his coinage of the HoH concept, Parfit (2011, p 616), speaks of "the next few centuries". It seems reasonable to assume that H(t) is not very volatile and spikey on time scales shorter than decades; otherwise we can replace it by a suitably smoothed approximation $H_{smooth}(t)$, which might for instance be defined as H(t) averaged over the interval [t, t +τ] for some suitable time window width τ such as τ = 100 years. In the following, unless otherwise specified, I will by "the HoH is now" mean "the HoH is in the present century".

Since we can affect the present and the future but not the past, it can be argued that a more action-relevant definition of HoH would be to consider only times from now and onwards when looking for the maximum of H(t). I will nevertheless stick to the convention of maximizing over *all* times, but will sometimes in what follows be lax about the possibility that the maximum is attained in the past.

To give a definition of H(t) that captures what we want and at the same time is precise enough to satisfy mathematical standards of rigor seems beyond what is presently doable, although there is likely scope for further work in this direction. The definition I will settle on is (almost) the one proposed by MacAskill (2022a) and is somewhat loose around the edges – although not so much so that it prevents a meaningful discussion of whether now is the HoH.

Here's a natural first idea. Assume that we have some utilitarian measure of how well things are going, and let V_∞ denote the total value of the world, from the beginning of prehistory until the end of time. Modeling the trajectory of human history as a stochastic process, this makes V_∞ a random variable. We write F_t as short for a complete description of everything that has happened up until time t, and V_t for the expected value $E[V_\infty | F_t]$ of the total value of the world given its evolution up to time t. Assuming V_∞ to be bounded guarantees existence of this expected value.[4] Now take τ to be a suitable time scale for what we take to be the duration of "now", such as τ = 100 years. By the tower property of conditional expectation (Williams, 1991) we have $E[V_{t+\tau} | F_t]=V_t$. In other words, given all

[3] While the present paper is largely a reply to MacAskill (2022a), it is worth noting that in his later book (MacAskill, 2022b), he looks more favorably upon urgent action, and my impression is that our actual disagreements at present time are less stark than what comes across in the following sections.

[4] This is reasonable if we can ignore everything that happens outside our future light-cone and after the heat death of the universe, but might, depending in part on unsettled cosmological issues, turn out untenable, causing deep difficulties that however fall outside of the ambitions of the present paper; see, e.g., Bostrom, 2011, and Askell, 2018.)

that has happened up to time t, the expected change over the time interval [t, t +τ] in the conditional expectation of V_∞ is precisely 0. Then the conditional variance $\mathrm{Var}[V_{t+\tau}|F_t]$ suggests itself as a candidate for a measure of how much is at stake at time t.

However, the problem with taking H(t) to be $\mathrm{Var}[V_{t+\tau}|F_t]$, from the point of view of the policy-relevant issue of patient vs urgent longtermism treated by MacAskill (2022a) and in Section 5 below, is that it fails to take into account the aspect of human agency. Imagine (counterfactually) that in the year 1200, a comet large enough to wipe out all large land-living animals including *Homo sapiens* were on collision course to impact the Earth in 1250, with the proviso that its passages near Jupiter and through the asteroid belt would involve chaotic dynamics making hit-or-miss a 50-50 chance issue in our world model. That would make $\mathrm{Var}[V_{1300}|F_{1200}]$ very large and the 13th century extremely hingey – perhaps enough to make it the true HoH – and yet it would not constitute a reason for 13th century decision-makers to take immediate longtermist action, because this was so long before reaching a technological level enabling a detection-and-deflection program that there was simply nothing anyone could do to affect the outcome.

So from the policy-relevance perspective, it would be desirable to pin down how much of $\mathrm{Var}[V_{t+\tau}|F_t]$ emanates from human decision-making between times t and t +τ. Instead of going down this path, however, I will follow more closely that of MacAskill (2022a), which is more explicitly tied to the policy issue at hand. He looks at how much a philanthropist at time t can improve V_t per amount of resources spent, and takes H(t) to be the optimal such ratio. The basic logic is simple: if now is time t, if t' > t, and if H(t') > H(t), then this suggests saving the resources until time t' and using them then as a way to achieve a greater improvement in V_t compared to using them now. (Of course, issues involving interest rates and financial risk greatly complicate this basic logic; see Trammell, 2021, for an in-depth treatment.)

This definition of H(t) is a stepping stone towards defining the HoH, and I do it slightly differently from MacAskill who looks at influentialness of individual people rather than hinginess of times, but the reasoning is essentially the same. He does however bring up some caveats and elaborations, two of which merit repeating here: one about (a) the definition of "spend", and the other about (b) what sort of spending counts as achievable. As regards (a), spending here only refers to direct actions towards improving V_t, as opposed to, e.g., investments or movement building aimed at attaining more resources for taking such action in the future. The possibility of grey areas (treated in more detail by Cotton-Barratt, 2020, and Mogensen, 2022) in this distinction does not preclude having a principled discussion about hinginess and the HoH.

Concerning (b), MacAskill emphasizes that we should only consider those actions as available at time t for which there is information available to philanthropists and other decision-makers at the time that gives plausible reasons to believe that the given action actually does improve V_t. This means that if I had been a citizen of Vienna in 1889 with an opportunity to kill baby Hitler, and if such an action would have had a great positive impact on political and other developments in the 20th century, that would still not count in this context, because at the time I would have had no way of knowing that the murder would be a good thing. I will not here address the topic of cluelessness (Greaves, 2016), which is the idea that perhaps the long-term effects of all our actions are similarly intractable, except to say that I am sympathetic to the stance of Greaves (2020) and Schubert (2022) that while the concept does bring up interesting concerns, there are still plenty of things we can do that make sense from the longtermist perspective of improving V_t.

Finally, when it comes to defining the HoH, I will do the straightforward thing of taking it to be the time t at which H(t) is maximized. This may at first look like it's in full agreement with MacAskill's understanding of the HoH-is-now hypothesis to mean that

> we are among the very most influential people ever, out of a truly astronomical number of people who will ever live,

but it isn't, due to the proviso "out of a truly astronomical…". If human history ends in extinction today, then the total number of humans who ever lived will stop at something like 10^{11}, which is not "a truly astronomical number", so that on MacAskill's definition there will be no HoH. MacAskill is explicit about this proviso and has a clear reason for his choice (which I will come back to in Section 3). Still the proviso is somewhat unnatural (and Mogensen, 2022, goes as far as saying that it "distorts the issue"), so I will drop it here. If H(t) is higher now than ever before, and we fail – in the above-quoted words by Tegmark – to "get our act together" but instead cause our near-term extinction, then I will say, contra MacAskill, that we did now live at the HoH, but we blew it. (Let's not!)

3. MacAskill's base rate argument against the Hinge of History being now

One of the most common arguments for longtermism involves back-of-the-envelope calculations in the style of Bostrom (2013) to show how much potential value there is in the future provided we avoid near-term extinction. Conservative assumptions about a kind of sustainable business-as-usual civilization refraining from space colonization lead Bostrom to suggest the potential for 10^{16} future human lives of normal (10^2 years) duration, while with interstellar and intergalactic space colonization 10^{32} lives may be feasible, and in case of a workable mind uploading technology, perhaps 10^{52}. The point here is not the exact orders of magnitude (which may well be somewhat off) but rather how very large these numbers are, in particular compared to numbers such as 10^{10} that a naïve short-termist thinker (or a proponent of so-called person-affecting views; see, e.g., Greaves, 2017) might otherwise be tempted to use to represent how many human lives are at stake in connection with existential risk.

For his main argument against the HoH-being-now hypothesis, MacAskill (2022a) combines such population estimates with another Bostrom idea, namely the formalization and quantification of the Copernican and mediocrity principles that has become known as the Self-Sampling Assumption (SSA), which Bostrom (2002) defines as follows:

> One should reason as if one were a random sample from the set of all observers in one's reference class. (p 57)

The choice of "reference class" may here depend on background information and the particular application at hand, and is often open to debate; in my case, the reference class could be, e.g., the set of all present-day Swedish-born professors of mathematics, or of all human beings throughout past and future history, or of all sentient beings who ever existed or will exist in the entire universe.

Now assume that one or the other of the future scenarios underlying Bostrom's population calculations is roughly what will happen, take $n = 10^{10}$ (slightly above the world population as of today), and let N be the total number of humans throughout past and future history. MacAskill notes that the ratio n/N will then be a very small number; exactly how small is less interesting, but he speaks about "one in a million trillion", so for concreteness let's say $n/N = 10^{-18}$. Imagine ranking the N humans that will ever live according to the peak value of hinginess H(t) that they experience during their lives (with some arbitrary tie-breaking convention), and note that applying SSA with reference class consisting of all N humans gives me probability n/N = 10^{-18} of being among the top n people on this ranking. If the HoH is now, then clearly I will be among these top n, so the probability that the HoH is now can be at most 10^{-18}.

MacAskill doesn't mean this to be an upper bound of the probability, given the best of our knowledge, that we are living at the HoH. Rather, it is the *base rate*, to be used as a Bayesian prior, and then modified by conditioning on all the evidence for or against living at the HoH that we get when we look around at the world in all its hinginess. Since 10^{-18} is such an extremely small number, the statement that "we live during the HoH" can be seen as quite an extraordinary claim, suitable for an application of the famous Carl Sagan maxim that "extraordinary claims require extraordinary evidence". And while MacAskill does admit that we do have evidence for living at the HoH, he holds evidence to be insufficiently extraordinary to bring the probability from 10^{-18} all the way up to (say) double-digit percentages. The probability that we live at the HoH is therefore small.

Note however that MacAskill's argument requires the premise of a large future a la Bostrom: it says nothing, for instance, to rule out the possibility of H(t) taking its maximum value now, promptly followed by an existential catastrophe wiping out humanity. Now we understand the inclusion of "a truly astronomical number" in his preferred definition of HoH: it allows him to state his main conclusion more crisply as a claim about a probability rather than a conditional probability. The crispness comes however at the expense of potential misleadingness to cursory readers. This is not, however, among my main issues – to be treated in the next section – with MacAskill's argument.

4. Counterarguments

I will offer two separate counterarguments to MacAskill's application of SSA for arguing against the hypothesis of the HoH being now. The first will be a general caution against giving too much credence to conclusions derived from the SSA principle and related anthropic arguments, and the second will be an observation that a prior probability of $1-10^{-18}$ is not as overwhelmingly unyielding to contrary evidence as first meets the eye.

Regarding how much credence to attach to SSA-based reasoning, we should first note that the principle is nowhere near as well-established and widely tested as other epistemic principles such as Bayesian updating in the light of new evidence or Ockham's razor. While SSA-like ideas go back a bit further than to Bostrom (2002) – see, e.g., Gott (1993) and Leslie (1998) – it remains the case that SSA and related principles have undergone very little practical testing. Applications of SSA are limited mainly to simple thought experiments and large-scale (often cosmological) questions whose answers remain unknown, with very little in between. If SSA amounted to a fairly obvious and uncontroversial claim, this in itself would perhaps not be a big problem, but that is not the case. On the contrary, there are several serious challenges regarding its validity, and while I do think SSA is an excellent idea that merits our attention, I also think that those challenges need to be taken seriously, including the following.

An obvious problem with SSA is the vagueness of what constitutes the correct reference class, but even if that can be settled, we must recognize that the idea that we can view ourselves as chosen from this class according to uniform distribution is a non-innocent probabilistic model assumption (Häggström, 2007). This model assumption needs better grounding, especially since no mechanism for it has been proposed. Such mechanisms would seem to require dualistic ideas along the lines of God picking each of us from some reservoir of souls and assign us to bodies chosen according to uniform distribution in the appropriate reference class – so naturally Bostrom and other defenders of SSA instead fall back on **the principle of indifference**, which was important in the early (pre-Kolmogorovian) development of probability theory, and states that in finite settings without obvious asymmetries between outcomes, equal credence should be assigned to all outcomes. Indeed, in the basic thought experiments used by Bostrom (2002), the reference classes have such far-reaching symmetries that it becomes difficult to defend a violation of the principle of indifference. In real-world applications such as MacAskill's, there is however an abundance of asymmetries. Why insist on uniform distribution on the class of observers, when probabilities could well depend on, say, observers' longevity,[5] their earliness in history, the energy turnover of their brains, their intelligence, or who knows what?

Another concern regarding the SSA is the mathematical impossibility, due to the fact that no probability measure on the set of natural numbers exists which assigns equal probability to all elements, of applying SSA in an infinite universe with infinite reference classes. Furthermore, in the finite case, applications of SSA tend to land in unpalatable consequences involving Doomsday arguments (Leslie, 1998; Bostrom, 2002; Häggström, 2016; Thomas, 2021), simulation hypotheses (Bostrom, 2003), and Boltzmann brains (Carroll, 2021). While I do have some sympathy for the intellectually courageous practice of biting the bullet and accepting wild consequences of otherwise reasonable-looking model assumptions, it still makes sense to take such consequences as additional evidence against the model assumptions if these are already seen to stand on shaky grounds.

All of this suggests that while an SSA approach to a problem like whether we live at the HoH provides a valuable perspective, the results of such an analysis need to be taken with a grain of salt. This is especially true when, as in the present case, a hypothesis which does have evidence pointing in its favor is assigned a very small probability in the SSA analysis.

Let me nevertheless, for the sake of argument and for concreteness, assume that MacAskill's analysis as outlined in Section 3 yields an appropriate prior for the HoH problem, and take the prior probability that we do *not* live at the HoH to be the seemingly overwhelming $1\text{-}10^{-18}$. What will be the consequences of this when we update the prior by conditioning on further information about our present situation?

It may be tempting to think of the prior probability $1\text{-}10^{-18}$ of not living at the HoH as so overwhelming, and of the extraordinariness of the evidence needed to overcome it as so great, that pretty much no evidence short of a miracle will do. That would be a mistake, however, as in the standard statistical setting with independent and identically distributed likelihood observations, ratios increase (or decrease, depending on one's point of view) exponentially with sample size.[6] Shaving an order of magnitude off an odds ratio takes a constant, and in many cases not particularly large, number of observations: to overcome 100 to 1 odds enough to equalize takes roughly just twice as large a sample as for overcoming 10 to 1 odds, and overcoming 10^{18} to 1 odds takes a sample size that exceeds the original one by a mere factor 18.

As an example consider coin tossing. Suppose that we have a coin whose heads-probability q is strictly speaking unknown, but whose fairness (q = 0.5) we have such strong prior reason to believe in that when we set up the a priori distribution to reflect these prior reasons fairly, we assign

[5] In fact, Bostrom (2002) himself goes on, via the concept of **observer moments**, to advocate a refinement of SSA where observers get probabilities proportional to their longevity.

[6] The mistake is understandable in view of today's scientific practice of designing experiments with small sample sizes meant to produce p-values around 0.05 or 0.001 being so dominant that we tend to forget about the exponential behavior of likelihood ratios as sample size increases.

probability 1-10^{-18} to the event that q = 0.5. The remaining probability 10^{-18} is smeared out uniformly on the interval where q ranges from 0 to 1. How many coin tosses with 75% of them heads would it take overcome such a dogmatic-looking prior? Perhaps surprisingly, the answer is not astronomical. A straightforward calculation shows that 1000 tosses with 750 heads vs 250 tails would be enough to totally turn the tables and produce a posterior distribution that places probability more than 1-10^{-35} on the coin being biased with q > 0.5.

This shows that overcoming an a priori distribution as heavily slanted as MacAskill's against a given proposition need not be as big a deal as it might first appear. Quantifying the available evidence that our time is the true HoH is, however, a delicate matter that I will not be able to resolve here, and a skeptic might argue that while we do have some evidence for living at the HoH, it is doubtful whether it amounts to 1000 independent bits of information. Nevertheless, I will point to some concrete reasons why it is not unreasonable to think it may be enough overcome the 10^{18} to 1 odds that MacAskill stacks against it.

Consider for instance the fact that we live within the 100-year period immediately following our first setting foot on a planetary body other than on our home planet. It is hardly implausible to suggest that such an event might be special enough for what happens the century following to determine the rough structure of our subsequent colonization of the universe. This is obviously not a water-tight argument for the HoH necessarily happening within a century from that pioneering first step of space exploration, but assigning a prior probability of 1 in 10 to such a concurrence does not strike me as crazy. If we can do that, then the success of Apollo program allows us to immediately cancel 17 of the 18 orders of magnitude contained in the odds ratio in MacAskill's prior. At that point, finding further evidence to overcome one more order of magnitude (or a few more, to take the probability of HoH being now past the even odds equilibrium and towards domination of the posterior distribution) no longer seems like such a daunting task.

Similarly, we may consider the evolution of intelligence. Humanity's astonishing trajectory from being a relatively unremarkable species a million years ago towards world domination has had relatively little to do with, say, our muscular strength or our physical endurance: what sets us apart from other species is almost entirely our intelligence. The step of beginning to outsource this unique resource to machines therefore seems like at least as significant a transition in human history as those first steps on the Moon, so assigning a probability of 1 in 10 to the HoH happening in the same century as the breakthrough in artificial intelligence (AI) seems reasonable, and that gives us another reason to cut 17 of the 18 orders of magnitude contained in MacAskill's odds against the HoH being now. (On the other hand, treating the breakthroughs in space exploration and in AI as statistically independent events seems wrong, and we cannot naively stack those 17 orders of magnitude gained from the latter on top of those 17 from the former to arrive at the probability of our time being the HoH being an impressive 1-$10^{18-17-17}$=1-10^{-16}.)

These examples are in the spirit of Shulman (2019), who holds forth the Moon landing and a few other similarly unique aspects of our time as indicative of the HoH. However, a skeptic such as MacAskill (2019b) may respond that these must be compared to other potentially great transitions in the future in order to carry much credence. The reason we think of the dawning of space exploration and of AI as so significant to human history might simply be myopia: we are smack in the middle of these transitions, and it's only natural that we overestimate the importance of the here and now. We might, at some time in the far future, look back on these events and view them as relatively small steps compared to, say, the initiation of a controlled merger of the Andromeda and Milky Way galaxies, or the rollout of a technology for harvesting energy from false vacuum decay, or something entirely beyond our present imagination. This is possible, but seems to me at most moderately likely.

We could also look for more generic clues to our place in history. Polishing on an analysis by Hanson (2009), Karnofsky (2021a) looks at global economic output (operationalized as inflation-adjusted global GDP, but the details do not matter much), and finds this:

> Let's say the world economy is currently getting 2% bigger each year. This implies that the economy would be doubling in size about every 35 years. If this holds up, then 8200 years from now, the economy would be about $3{\cdot}10^{70}$ times its current size. There are likely fewer than 10^{70} atoms in our galaxy, which we would not be able to travel beyond within the 8200-year time frame. So if the economy were $3{\cdot}10^{70}$ times as big as today's, and could only make use of 10^{70} (or fewer) atoms, we'd need to be sustaining multiple economies as big as today's entire world economy *per atom*. (Emphasis in original.)

The last conclusion can be treated as a reductio, meaning that until the time we begin colonizing other galaxies, there will not be as many as 82 future centuries exceeding ours in relative GDP growth. Similar exercises with GDP replaced by energy consumption or population of flesh-and-blood humans lead to similar results; see Murphy (2011) and Abell (1982), respectively, with the latter culminating in the vivid image of a sustained 2% annual population increase leading within 5300 years to “a great sphere of humanity 150 light years in radius [which] would be expanding at its surface at the speed of light” (p 594). Sticking for concreteness to Karnofsky’s calculation, one might object that a greater number of centuries with such GDP increase can be achieved if GDP is allowed to oscillate, but the argument holds up against this objection by replacing momentary GDP by all-time-high GDP in the analysis. Also, if we replace 10^{70} with the upper bound of 10^{82} on the number of atoms in the (reachable) universe, we get at most 95 future centuries of present-day GDP growth or more, even without the restriction of staying in the Milky Way. Plausibly, times with unusual amount of change also come with unusual hinginess, so it seems reasonable to assume that the probability of the HoH happening in one of the 95 centuries exhibiting the most drastic economic growth is at least 1/2. Conditioning on the fact that we live in such a century we arrive at a probability of $1/(2{\cdot}95) = 0.0053$ that the HoH is in the present century, thereby overcoming nearly 16 of the 18 orders of magnitude in MacAskill’s prior.

Another circumstance that points towards the probability of our living at the HoH being far larger than 10^{-18} are how very early in human history we find ourselves in relation to the hugeness of the Bostromian futures that do most of the work in MacAskill’s argument: it is not unreasonable to think that the long-term fate of humanity is disproportionally determined by what happens very early. In the comments section to MacAskill (2019a), such considerations led Ord (2019) to suggest a Laplacean $1/t^2$ prior for the timing of the HoH, and “a prior chance of [the present being] HoH of about 5% or 2.5%”.

Relatedly, our living at a time when human population counts in billions rather than the quadrillions or more as in those Bostromian futures (aided, perhaps, by space colonization and/or mind uploading) presumably makes it easier for individuals and small groups to have momentous influence on the remainder of human history, again pointing towards our time exhibiting a greater-than-usual hinginess and thereby a larger probability of being the HoH. Rather than plunging into (shaky) attempts to quantify the evidential value of these two circumstances, let me move on to yet another one, namely the existential risk that is gestured at in statements like the Tegmark quote in Section 1.

Assume for the moment that during the present century we run a 10% risk of a catastrophe resulting in the extinction of the human species. How many similarly dangerous (or worse) centuries can there be? For there to be n such centuries, we need to survive the first n-1 of them, which has probability at most $(1-0.1)^{n-1}$. For instance, taking n = 50 yields probability at most $(1-0.1)^{49} = 0.0057$ of having at least 50 centuries with such a high extinction risk. It seems plausible to postulate at least even odds for the HoH to happen during one of the 50 most dangerous centuries in terms of extinction risk, which would put a lower bound on the probability of the HoH happening during the present century at just under 0.5/50 = 0.01. Again, most of the orders of magnitude in MacAskill's odds against the HoH being now are wiped out.

But how realistic is it to postulate such a 10% risk of extinction catastrophe? For natural risks such as supervolcanoes and asteroid impacts, we do have a reasonable grip on estimating the risk level, and on time scales of a century they are small. Estimating anthropogenic risk is harder, as circumstances here are so new and rapidly changing that no scientific consensus exists that pinpoints even the appropriate order of magnitude of the total risk, but this is where the risk must come from in order to come anywhere near the suggested 10%. The magnum opus on existential risk (x-risk) by Ord (2020) estimates, along with suitable declarations of epistemic modesty, the x-risk probability during the next 100 years as 1/6. This, however, is not immediately applicable to the calculation in the preceding paragraph, because the concept of x-risk does not quite equal that of extinction risk: Ord follows Bostrom's (2013) standard definition of x-risk which besides extinction risk includes other events on a similar level of badness in terms of losing humanity's potential for future flourishing.

To bridge the gap between x-risk and extinction risk in our calculation we can choose either of two approaches. One is to note that on a centennial time scale, and judging from Ord (2020) along with other works on existential risk including Bostrom and Cirkovic (2008), Pamlin and Armstrong (2015), Häggström (2016) and Yudkowsky (2022), extinction risk seems to be the main and most likely example of x-risk, so we could take Ord's 1/6 and round it down to something like 1/10 to get a plausible extinction risk probability. The other approach is to apply the calculation to x-risk rather than extinction risk, which however gives rise to the complication that unlike extinction events, existential catastrophes can in principle happen more than once: the first time around, humanity survives but loses most of its potential future value; the second time, most of the remaining potential value is lost, and so on. This can arguably be fixed by noting that the HoH will likely happen before (or during) the first existential catastrophe, because after that most of the potential value is already out of reach, so it makes sense to consider only those centuries that predate the first existential catastrophe, and with that restriction the calculation goes through as for extinction risk.

As mentioned, there is considerable scope for debate regarding Ord's (2020) assignment of probability 1/6 to x-risk during the next 100 years, but it does not strike me as an obvious overestimate. Regarding the specific risk source that he gives the largest probability – unaligned AI, which accounts for probability 1/10 – it seems to me that on the contrary, the best state-of-the-art judgements point towards, if anything, the risk being even greater. See, e.g., Ngo (2020), Häggström (2021), Carlsmith (2022), Christiano (2022) and Cotra (2022a,b) for recent discussions of the gravity of AI x-risk, and Yudkowsky (2022) for a particularly alarming but not obviously erroneous account of the difficulty of managing that risk.[7]

Still, in view of the large uncertainties, it is not clear how much weight to give x-risk in the HoH discussion. As a stand-alone argument for our living at the HoH, I find it relatively strong, but as an

[7] See also Finnveden, Riedel and Shulman (2022) for other AI-related reasons (beyond AI x-risk) for our time being the HoH.

objection to MacAskill's (2022a) base rate argument it is open to the counter that Ord's 1/6 estimate has not taken the base rate argument into account and might therefore need to be adjusted down.

Be that as it may, we have in the above seen a wide array of circumstances that make our time unusual in a way that should boost our credence it being the HoH. I have not provided any mathematical model for integrating this evidence and deriving a Bayesian posterior, which would be a daunting task and probably not achieve much beyond a false air of precision, but I can offer my subjective judgement of where the evidence seems to take us: a reasonably high probability of our living at the HoH, most likely a double-digit percentage number, and plausibly even the majority of the probability mass.

However, as good Bayesians we cannot be content with looking for evidence in favor of a given proposition. Rather, we should aspire to take into account all relevant information, including that which points in the opposite direction. MacAskill (2022a) suggests the following circumstance as evidence against the HoH being now: our ability to affect the future for the better depends on our knowledge about how the world works, and this (collective) knowledge has improved in the past and can be expected to improve further in the future, pushing our ability and therefore the level of hinginess in the same direction, thus improving the chance that the HoH is in the future rather than now. This is an interesting argument, but it is unclear whether it works, because it may be that the most relevant quantity, given the definition of hinginess H(t) recalled in Section 2, is not the absolute level of knowledge but rather the knowledge gap between on one hand philantropists aiming to increase V_t, and on the other hand decision-makers in general; the absolute level increasing over time need not imply that gap does. I will return briefly to this issue in Section 5.

Relatedly, MacAskill suggests that the anomalous GDP growth (along with other rapid societal changes) discussed above might not tend to increase hinginess but rather to decrease it, because it might be easier for V_t-maximizing philantropists to influence the world in periods with little change when they are in a better position to predict the consequences of their actions, and are therefore better able to choose the best actions. This could be, but one can also reason in the opposite direction and argue that in times of technological and societal stasis it is very difficult to have much impact, and that the best opportunities for impact are when society opens up to change. One can argue back and forth over which of these effects is the stronger. I believe more in the latter, but even if we grant MacAskill even odds on this issue, the effect on the probability distribution of hinginess of our present turbulent time (relative to others) will mostly be to push probability from the middle and towards the extremes, resulting in an increased probability that we live at the HoH.

5. In defense of urgent longtermism

Recall from Section 1 distinction between the urgent and patient longtermism, and how the choice between them relates to the HoH issue. The arguments offered in the previous section for the likelihood of our living at the HoH point somewhat in the direction of favoring an urgent approach, while still falling far short of settling the debate, mainly for two reasons. First, while I have argued that it is reasonably likely that the HoH is now, this does not entirely eliminate the possibility that the HoH lies in the (possibly far) future. Second, even if the HoH is now, there could be other times in the future that also exhibit high hinginess, and by saving in financially cunning ways that give good return on investment and delaying action to those later times, we might create more expected value. So more can be said on the choice of patient vs urgent longtermism, and the following are a few thoughts on this matter.

Consider for concreteness a scenario where our present century is fairly hingey, but not quite as hingey as the true HoH that takes place a million years from now. The choice of time scale here may look like a straw man of MacAskill's (2022a) position, but this sort of very large time scale is in fact needed in his base rate argument for a very small prior probability of the HoH being now. For further concreteness, add the assumption that from now on and over the next million years, we are able to raise \$1Bn (one billion dollars) per year for longtermist purposes. Ignoring for the moment the issue of interest rates, we may note the asymmetry that, among the full $\$10^6$Bn raised over this period, only at most \$100Bn can be spent during the present century, while potentially the entire $\$10^6$Bn is available at the HoH at the end of this period. Thus, in this idealized situation, the patient vs urgent longtermism issue boils down to whether to spend \$100Bn this century and the remaining \$999,900Bn (i.e., about ten times present-day annual GDP) during the HoH a million years hence, or to save the entire amount until then.

It is by now a well-established wisdom in the effective altruism community that for individual donors, the phenomenon of diminishing marginal returns is usually negligible, so the effect of spending on a given cause can be seen as linear and one might as well concentrate all one's spending on one single cause, but that on larger scales diminishing returns may be more substantial (see, e.g., Kuhn, 2014, and Snowden, 2019). Whether in the situation outlined here \$100Bn in the present century has more impact than the additional 0.01% of longtermist funds at the distant HoH is of course an open question, but it seems at least plausible that diminishing returns would cause the urgent spending to be the more impactful option.

Next, consider the effect of interest rates. Even a modest real interest rate of 1% per year would mean a doubling of capital in 70 years, and a multiplication by 2.7 each century. So the value a million years down the road of the \$100Bn raised in the present century will be \$100Bn times $2.7^{10,000}$, which means, with a=2.7, that the present century's fraction of that raised in the next million years will be

$$1/(1+a^{-1}+a\text{-}^{2}+\ldots+a^{-10,000}) = (1-a^{-1})/(1-a^{-10,001}) = 0.63.$$

This is a very different story from that obtained with zero interest rate, but $100\text{Bn}\cdot 2.7^{10,000}$ exceeds 10^{4000} – which in turn far exceeds the number of elementary particles in the visible universe or pretty much any (non-combinatorial) number that arises in our world, so we see that the result is nonsensical for essentially the same reason as in the GDP and population growth examples in Section 4 (i.e., sustained exponential growth quickly becomes incompatible with finiteness of the speed of light). In conclusion, working with constant interest rates over such enormous time spans is infeasible. Grabbing for straws, perhaps something like hyperbolic discounting (Ainslie, 2001) could come to the rescue, but I think the grim truth is that we have at present no way of making sense of the value of financial assets over millennial or longer time scales.

We could, however, point to some more concrete issues concerning such extremely long-term investments. One is that the concept of money is just a few millennia old, and modern capitalism only a couple of centuries, so that saving over time scales longer than that implies a faith in the longevity of these institutions that appears unwarranted. Who knows what kinds of governmental confiscations, wars, revolutions and civilizational collapses the next one million years may have in store for us?

On the more positive side, consider moral progress, of which recent centuries have seen plenty, as MacAskill (2022a) points out in connection with the argument about the trend towards improved knowledge discussed in Section 4 above:

> In 1600 [it was] believed that women and people of other races and religions are of lesser moral standing than European Christian men. Intense social hierarchy, inequality and slavery were regarded as the natural and just way of things. Homosexuality and premarital sex were regarded as deeply immoral. The idea of liberalism had not been developed. Torture was commonplace and celebrated, as were cruel punishment and violence against heretics.

These views and practices are now understood as morally wrong, and MacAskill goes on to stress that the trend continues and that in the last 50 to 100 years "rights for women, minorities and people of all sexual identities have been progressively secured". His point here (echoing the famous final paragraph of Parfit, 1984) is that this development towards better morality can be expected to continue in the future, and I agree, but I think the consequence of this may go beyond what MacAskill suggests about future longtermists being better equipped to know what to do. Consider this:

Assuming that saving in longtermist funds does not influence GDP growth in any particular direction, the point of these savings is to exert influence on how society in the future will use their resources, by earmarking some of these for longtermist goals. If human morality continues to improve, and if effective altruism and V_t-maximization are largely correct stances (otherwise the whole patient vs urgent longtermism issue seems moot, and it is probably not a good idea constrain future generations to act unethically), then we can expect society as a whole to eventually adopt these ethical views and act upon them, thereby removing the need for us to lock up resources in longtermist funds in order to constrain future people to act on these views.

In practice, yet another concern one may have regarding patient longtermism is the following. While near-term and long-term goals for society are often fairly well aligned, there will be cases when they come somewhat apart, and it is often difficult in such cases to convince individuals and society as a whole to sufficiently prioritize the long term (as evidenced by, e.g., climate debate). It is likely that this difficulty is exacerbated if the long-term action is to put money in the bank rather than, say, taking concrete action against some x-risk.

The discussion so far points mostly in the direction of preferring urgent longtermism, but it is interesting to stop and consider how general the arguments are. If they suggest that taking concrete action is better than patiently saving now in the 21st century, will the conclusion be the same in, say, the 22nd century, or the 24th? Are the arguments in fact so general as to suggest that we should *never* adopt a patient longtermist stance? Well, while some of the arguments are fairly general, I'd say the largest cornerstone in the case for urgent longtermism presented so far consists of the evidence discussed in Section 4 for the unusually large hinginess of the present century. If this (apparent) hinginess level goes down in later centuries, the case for urgent approaches will then become weaker.

One can turn this question around and ask whether, for someone who accepts MacAskill's base rate argument, we can ever expect to see evidence of being at the HoH that is so much stronger than today's already quite remarkable evidence in this direction that it overcomes the low base rate and recommends urgent action? Maybe, but then again maybe not, and it would be sad to see longtermist funds remaining untouched forever.

6. Concluding remarks

Despite all the evidence presented in Section 4 for our living at the HoH, the question for whether or not that is the case remains wide open. Likewise, the discussion in Section 5 hardly settles the issue of patient vs urgent longtermism. While I do think the idea of setting up longtermist funds with resources to be harvested in a million years can (at least for the time being) be written off as impractical and overly fanciful, there is the much more modest idea that the 22nd or the 24th century might be even hingier than the 21st, and that a good longtermist move would be to save resources until then. The arguments in the present paper do relatively little to invalidate such a moderate version of patient longtermism. On the other hand, such a view receives hardly any support from MacAskill's base rate argument, which needs much longer time scales to produce the seemingly overwhelming prior odds that he holds forth against our living at the HoH.

Regarding the choice between urgent action and this more moderately patient longtermism, there is the following basic intuition. While it remains a highly open question which of these would be recommended by a well-informed utilitarian-style calculation, there is an asymmetry in that we pretty much know that the level of existential risk in the present century is alarmingly high (in particular, it has become increasingly clear in the last couple of decades that AI existential risk is substantial). If we fail to take concrete action in mitigating this risk, we may face an existential catastrophe that cannot be repaired no matter the amount of resources put aside for longtermist work in later centuries. I think this intuition makes sense, and it is lent some support by the arguments in the present paper that urgent longtermism is at least not obviously wrong.

Acknowledgement. I am grateful to Gustav Alexandrie, Björn Bengtsson, Hilary Greaves, Erich Grunewald, Julia Jansson, Stefan Schubert, Christopher Star, Markus Stoor, Marcus Widengren and Anna Wisakanto for discussions and helpful remarks.

References

Abell, G. (1982), *Exploration of the Universe, 4th edition* (Thomson Learning, London).

Ainslie, G. (2001), *Breakdown of Will* (Cambridge University Press, Cambridge).

Askell, A. (2018), *Pareto Principles in Infinite Ethics*, Ph.D. thesis, New York University.

Bostrom, N. (2002), *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (Routledge, New York).

Bostrom, N. (2003), 'Are you living in a computer simulation?', in *Philosophical Quarterly* 53: 243-255.

Bostrom, N. (2011), 'Infinite ethics', in *Analysis and Metaphysics* 10: 9-59.

Bostrom, N. (2013), 'Existential risk prevention as global priority', in *Global Policy* 4: 15-31.

Bostrom, N. and Cirkovic, M. (2008), *Global Catastrophic Risks* (Oxford University Press, Oxford).

Carlsmith, J. (2022), 'Is power-seeking AI an existential risk?', https://arxiv.org/abs/2206.13353

Carroll, S.M. (2021), 'Why Boltzmann brains are bad', in S. Dasgupts, R. Dotan and B.Weslake (eds), *Current Controversies in Philosophy of Science* (Routledge, New York), 7-20.

Christiano, P. (2022), ‘Where I agree and disagree with Eliezer’, *LessWrong*, June 19.

Cotra, A. (2022a), ‘Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover’, *LessWrong*, July 18.

Cotra, A. (2022b), ‘Two-year update on my personal AI timelines’, *LessWrong*, August 2.

Cotton-Barratt, O. (2020), ‘Patient vs urgent longtermism has little direct bearing on giving now vs later’, *Effective Altruism Forum*, December 9.

Finnveden, L., Riedel, J. and Shulman, C. (2022), ‘Artificial general intelligence and lock-in’, https://docs.google.com/document/d/1mkLFhxixWdT5peJHq4rfFzq4QbHyfZtANH1nou68q88

Gott, R. (1993), ‘Implications of the Copernican principle for our future prospects’, in *Nature* 363: 315-319.

Greaves, H. (2016), ‘Cluelessness’, in *Proceedings of the Aristotelean Society* 116: 311-339.

Greaves, H. (2017), ‘Population axiology’, in *Philosophy Compass* 12: e12442.

Greaves, H. (2020), ‘Evidence, cluelessness and the long term’, talk at the virtual *Effective Altruism Student Summit*, October 24-25, with transcript at the *Effective Altruism Forum*, November 1.

Häggström, O. (2007), ‘Uniform distribution is a model assumption’, http://www.math.chalmers.se/~olleh/reply_to_Dembski.pdf

Häggström, O. (2016), *Here Be Dragons: Science, Technology and the Future of Humanity* (Oxford University Press, Oxford).

Häggström, O. (2021), *Tänkande maskiner: Den artificiella intelligensens genombrott* (Fri Tanke, Stockholm).

Hanson, R. (2009), ‘Limit to growth’, *Overcoming Bias*, September 21.

Harris, S., Goldstein, R. and Tegmark, M. (2018), ‘What is and what matters’, *Making Sense Podcast*, March 19.

Karnofsky, H. (2021a), ‘This can’t go on’, *Cold Takes*, August 3.

Karnofsky, H. (2021b), ‘The most important century (in a nutshell)’, *Cold Takes*, September 23.

Kuhn, B. (2014), ‘How many causes should you give to?’, https://www.benkuhn.net/how-many-causes/

Leslie, J. (1998), *The End of the World: The Science and Ethics of Human Extinction* (Routledge, New York).

MacAskill, W. (2019a), ‘Are we living at the most influential time in history?’, *Effective Altruism Forum*, September 3.

MacAskill, W. (2019b), Comment ‘I don’t think I agree with this, unless…’ on MacAskill (2019a), *Effective Altruism Forum*, September 13.

MacAskill, W. (2022a), ‘Are we living at the hinge of history?’, in J. McMahan et al (eds) *Ethics and Existence: The Legacy of Derek Parfit* (Oxford University Press, Oxford), 331-357.

MacAskill, W. (2022b), *What We Owe the Future* (Basic Books, New York).

Mogensen, A. (2022), ‘The hinge of history hypothesis: reply to MacAskill’, GPI Working Paper No. 9-2022, Global Priorities Institute, Oxford.

Murphy, T. (2011), ‘Galactic-scale energy’, *Do the Math*, July 12.

Ngo. R. (2020), ‘AGI safety form first principles’, https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view

Ord, T. (2019), Comment ‘Hi Will, It is great to see…’ on MacAskill (2019a), *Effective Altruism Forum*, September 6.

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury, London).

Pamlin, D. and Armstrong, S. (2015), *12 Risks That Threaten Human Civilization* (Global Challenges Foundation, Stockholm).

Parfit, D. (1984), *Reasons and Persons* (Clarendon, Oxford).

Parfit, D. (2011), *On What Matters, Volume 2* (Oxford University Press, Oxford).

Schubert, S. (2022), ‘Against cluelessness: pockets of predictability’, blog post, May 18, https://stefanfschubert.com/blog/2022/5/18/against-cluelessness-pockets-of-predictability

Shulman, C. (2019), Comment ‘I think this point is even stronger…’ on MacAskill (2019a), *Effective Altruism Forum*, September 7.

Snowden, J. (2019), ‘Should we give to more than one charity?’, Greaves, H. and Pummer, T. (eds) *Effective Altruism: Philosophical Issues* (Oxford University Press, Oxford), 69-79.

Thomas, T. (2021), ‘Doomsday and objective chance’, GPI Working Paper No. 8-2021, Global Priorities Institute, Oxford.

Trammell, P. (2021), ‘Dynamic public good provision under time preference heterogeneity: theory and applications to philanthropy’, GPI Working Paper No. 9-2021, Global Priorities Institute, Oxford.

Williams, D. (1991), *Probability with Martingales* (Cambridge University Press, Cambridge, UK).

Yudkowsky, E. (2022), ‘AGI ruin: a list of lethalities’, *LessWrong*, June 6.