The hinge of history hypothesis: Reply to MacAskill

Andreas Mogensen (Global Priorities Institute)

Global Priorities Institute | August 2022

GPI Working Paper No. 9 - 2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

The Hinge of History Hypothesis: reply to MacAskill

Abstract: Some believe that the current era is uniquely important with respect to how well the rest of human history goes. Following Parfit, call this the Hinge of History Hypothesis. Recently, MacAskill has argued that our era is actually very unlikely to be especially influential in the way asserted by the Hinge of History Hypothesis. I respond to MacAskill, pointing to important unresolved ambiguities in his proposed definition of what it means for a time to be influential and criticizing the two arguments used to cast doubt on the claim that the current era is a uniquely important moment in human history.

1. Introduction

Some believe that the current era is a uniquely important moment in human history. We are living, they claim, at a time of unprecedented risk, heralded by the advent of nuclear weapons and other world-shaping technologies. Only by responding wisely to the anthropogenic risks we now face can we survive into the future and fulfil our potential as a species (Sagan 1994; Parfit 2011, Bostrom 2014, Ord 2020).

Following Parfit (2011), call the hypothesis that we live at such a uniquely important time *the Hinge of History Hypothesis* (3H). Recently, MacAskill (2022) has argued that 3H is "quite unlikely to be true." (332) He interprets 3H as the claim that "[w]e are among the very most influential people ever, out of a truly astronomical number of people who will ever live" (339) and defines a period of time as influential in proportion to "*how much expected good one can do with the direct expenditure* (*rather than investment*) *of a unit of resources at* [that] *time*" (335), where 'investment' may refer "to both financial investment, and to using one's time to grow the number of people who are also impartial altruists." (335 n.13) MacAskill thus relates the

truth or falsity of 3H to the practical question of the optimal time at which to expend resources to achieve morally good outcomes, considered impartially.

MacAskill presents two arguments against 3H. The first is an argument that the prior probability that we are living at the most influential time in history should be very low, because we should reason as if we represent a random sample from observers in our reference class. The second is an inductive argument that we should expect future people to have more influence over human history because the overall trend throughout human history is for later generations to be more influential.

In my view, neither of these arguments should convince us. As I argue in section 2, MacAskill's priors argument relies on formulating 3H in a way that does not conform to how this hypothesis is traditionally understood. Moreover, I will argue in section 3 that MacAskill's definition of what it means for a time to be influential leaves too many unresolved ambiguities for his inductive argument to work.

2. MacAskill's Priors Argument

In MacAskill's preferred formulation, 3H states that we are "among the very most influential people ever, out of a truly astronomical number of people who will ever live." (339) The population is assumed to be 'astronomical' because MacAskill builds in the assumption that there will exist very many people in future and those people might be distributed across many star systems. In his conception, 3H states that "not merely are we among the most influential people ever, but we are among the most influential people ever out of a civilization that will one day take to the stars." (339)

As noted previously, many of those who suppose that the current era is uniquely important do so specifically because they believe we live in a time of heightened extinction risk. This is the picture painted by Parfit (2011: 616) in illustrating the claim that we live at the 'hinge of history.' In his conception, this is the "most dangerous and decisive period" in our species' history, to be survived only "[i]f we act wisely in the next few centuries".

There exists, then, a mismatch between MacAskill's formulation of 3H and Parfit's conception. In MacAskill's formulation, the claim that we live at the 'hinge of history' builds in the assumption that humanity will continue to exist for long enough to establish a significant presence throughout the galaxy. In Parfit's conception, what characterizes the 'hinge of history' is that the continued existence of our species is especially uncertain. On MacAskill's interpretation, if we knew we lived at the 'hinge of history,' no urgency whatsoever would attach to lowering the risk of extinction this century, since 3H entails that humanity will not go extinct for a very long time. In Parfit's conception, the 'hinge of history' is a time at which lowering the risk of human extinction is especially urgent.

This, to me, suggests that MacAskill's formulation distorts the issue – or at least changes the subject. But perhaps this can be easily fixed. One way to remove the inconsistency would be to modify MacAskill's formulation so that it states not that the *true* future population is astronomical in size, but rather that the *expectation* of the total future population is astronomical in size. Since the claim that the expectation of the total future population is very large can be accepted by someone who assigns significant probability to near-term extinction, the mismatch between MacAskill's formulation of 3H and the Parfitian conception of what it means to live at the 'hinge of history' would be dissolved.

However, MacAskill's first argument against 3H cannot withstand this change in formulation. To see why, let me set out the argument. It appeals to a principle governing self-

locating beliefs due to Bostrom (2002), known as the *Self-Sampling Assumption* (SSA). Fix some property F. Let N be a random variable whose value is the number of observers in my reference class. Let M be a random variable whose value is the number of those observers who exhibit F. Let '$F\alpha$' denote the proposition that I exhibit F. Then SSA states that my prior should be such that

$$\Pr(F\alpha \mid M/N = m/n) = m/n$$

It follows that if I update by conditionalizing on my evidence, learning that m out of n observers in my reference class exhibit F will lead me to set my credence that I exhibit F to be m/n. If F is a superlative property, such as being the strongest or the most influential, m/n necessarily equals $1/n$.

MacAskill argues that SSA entails that the unconditional prior probability that we are among the most influential people to live throughout what remains of human history should be extremely low. He writes that

> there are plausibly a vast number of people in the future. … [T]here are one hundred billion stars in the Milky way; settling just 0.1% of them with the same population as on Earth would mean that there are a trillion trillion people to come. … If there are a trillion trillion people to come, then the a priori probability that we are among the million most influential people ever is one in a million trillion. (340)[1]

[1] This argument is anticipated and discussed by Sagan (1994: 306-8). Roughly, Sagan argues that our evidence that the present time is extraordinary is strong enough to sufficiently raise the very low prior probability of 3H. See MacAskill (2022: 341-4) for the contrary argument that our evidence is too

This line of argument assumes MacAskill's chosen formulation of 3H, since it assumes that there will be trillions of people to come. In the Parfitian conception, if we live during the 'hinge of history,' we are *not* entitled to assume that there are trillions of people to come. Suppose that we replace the claim that we *will* be succeeded by very many descendant generations with the claim that the *expected* future population is astronomical in size. It may be tempting to suppose that SSA entails that upon learning that the expected future population is astronomical in size, we should also set our credence that we are among the million most influential people to be very low. For example, we might suppose that if N is a random variable that denotes the size of the future population numbered in units of a million persons and $\mathbb{E}(N)$ is its expectation, then the prior probability that we are among the million most influential persons should be $1/\mathbb{E}(N)$. Therefore, if $\mathbb{E}(N)$ is enormous, the prior probability that we live at the 'hinge of history' should be miniscule.

This line of reasoning is mistaken.[2] Given SSA, the prior probability that we are among the million most influential persons should not be $1/\mathbb{E}(N)$, but $\mathbb{E}(1/N)$. Where F is the property of being among the million most influential people still to come, SSA entails that a rational agent's prior conditional probability obeys

$$\Pr(F\alpha \mid N = n) = 1/n$$

impoverished to sufficiently raise the posterior probability of 3H given the very low prior suggested here.

[2] Similar points are raised by William Kiely in online discussion of MacAskill's argument. See MacAskill (2019).

Then, by the Law of Total Probability,

$$\Pr(F\alpha) = \sum_n \Pr(F\alpha \mid N = n)\Pr(N = n)$$

$$= \sum_n (1/n)\Pr(N = n) = \mathbb{E}(1/N)$$

Moreover, there is no reason in general to assume that $1/\mathbb{E}(N)$ and $\mathbb{E}(N)$ are approximately equal. They especially come apart in cases where low values for N are not too unlikely, and so play a dominant role in determining the value of $\mathbb{E}(1/N)$, whereas high values for N are not too unlikely, and so play a dominant role in determining the value of $\mathbb{E}(N)$.[3] This is not too unlike the situation in which we find ourselves if we live at the Parfitian 'hinge of history.'

MacAskill's argument might still go through if we look not to the future, but to the past. By a conservative estimate, there have existed at least 100 billion human beings (Curtin 2007). By SSA, the prior probability that we are among the million most influential people who have ever existed given that there have existed at least 100 billion human beings is at most .00001. Assume that our epistemic standing with respect to the stated estimate of the total population to date is good enough that it forms part of our evidence. Then the prior probability that we are among the million most influential people who have ever existed should be no greater than .00001.

[3] For example, if $\Pr(N = 1) = 0.1$, $\Pr(N = 10) = 0.8$, and $\Pr(N = 1{,}000) = 0.1$, then $\mathbb{E}(1/N) = 0.1801$ whereas $1/\mathbb{E}(N)$ is nearly twenty times smaller at 0.00925.

The claim that we live at the 'hinge of history' arguably does build in a comparison between present and past. When Ord (2020: 19-23) makes the case that we live at a uniquely influential point in human history, he primarily emphasizes the uniqueness of the risks we face relative to our forebears, especially the claim that the advent of nuclear weapons represents the first time in history that anthropogenic risks to human survival exceed natural risks. Nonetheless, the backward-looking priors argument does not help MacAskill's case, for two reasons.

The first reflects the practical orientation of MacAskill's discussion. For the question of whether to expend our resources now or invest them, what ultimately matters is how our influence compares to that of people who will exist in future. How our influence compares to that of past people matters only insofar as it may provide evidence about how our influence compares to that of our descendants. The decision-relevant question is whether we are among the very most influential people out of all those people who now exist or will exist in future.

The second reason why the backward-looking argument does not help MacAskill's case is that he maintains that the current time is more influential than any previous time. This claim plays an important role in his second argument against 3H.

3. MacAskill's Inductive Argument

MacAskill's second argument against 3H is inductive. According to the inductive argument, the influence of comparable people in the past has been increasing over time, particularly as a result of gains in knowledge. Absent contrary evidence, it should therefore be expected to continue increasing within the foreseeable future. Therefore, the current era is probably not the most influential of those remaining in our species' history.

To support the claim that influence has been increasing over time, MacAskill asks us to compare ourselves to well-educated Europeans living in 1600. He argues that their opportunities to shape the long-run future were meagre by comparison with ours today, with the possible exception of their ability to bring about persistent changes in values. In large part, he notes, that is because of their impoverished scientific understanding: "They could not have known about the vastness of the future nor make reasonable guesses about how to positively influence the long-run future." (347) Most importantly, MacAskill notes, their moral beliefs seem to us badly wrong in many different ways, "grounded in a narrow understanding of Christian doctrine that we would now deplore." (347) MacAskill claims, moreover, that if we make similar comparisons between the current time and dates in the more recent past, such as 1920 or even 1970, we find that "there is a good argument for thinking that we are in a much better position to have a positive impact today than we could during those times." (347)

There are at least two reasons why we should be sceptical of this argument from the outset.

The first I have already noted. MacAskill's argument predicts that we now live at the most influential time relative to all previous historical eras. In fact, the inductive argument seems to use this claim as one of its key intermediary conclusions. MacAskill (2022: 346-7) devotes considerable space to arguing that we are better placed to have a positive impact than Europeans alive during the Early Modern period, and he extends this argument to times as recent as 1970. He says little to justify thinking that people at earlier times were, in general, more influential than people at still earlier times. The argument is driven primarily by the claim that we are unusually influential relative to earlier times. Given SSA, we should assign a low prior probability to that hypothesis.

Moreover, the claim that people's influence over human history has been increasing over time is surprising because of the temporal asymmetry of causation. Who controls the past controls the future. How could we be more influential than past people?

The answer, I take it, is that this can be so once we fix a technical definition of 'influential' like that proposed by MacAskill. Recall that he defines a time as influential in proportion to "*how much expected good one can do with the direct expenditure* (*rather than investment*) *of a unit of resources at* [that] *time*" (335), where 'investment' may refer "to both financial investment, and to using one's time to grow the number of people who are also impartial altruists." (335 n.13) Insofar as the impacts achievable by past people run via their influence on the opportunities, capabilities, and goals of present people, the relevant outcomes might be said to fall within the scope of good achievable via investing, rather than direct expenditure.

The plausibility of MacAskill's inductive argument rests, therefore, to a large extent on how we define 'influential' and especially how we understand the distinction between direct expenditure and investment. I claim that the proposed definition is subject to significant unresolved ambiguities and deployed at points in contradictory ways. For this reason, I claim, it cannot bear the weight it is required to bear for the argument to come out cogent.

First and foremost, significant unclarities attach to the distinction between direct expenditure and investment. As suggested above, this distinction is handled in apparently contradictory ways. Recall that the good that can be achieved through investing at a time does not contribute to how influential that time is. What matters is the good that can be achieved through direct expenditure. Recall, moreover, that investment refers "to both financial investment, and to using one's time to grow the number of people who are also impartial altruists." (335 n.13) Recall also that in arguing that past people were less influential than we

are, MacAskill claims that their opportunities were less high-leverage than ours today, but notes, as a possible exception, "the opportunity to shape the values of the time, which are plausibly persistent for a long time period, including via religious institutions." (347 n. 32) This suggests that expending resources so as to achieve persistent changes in values counts as a form of direct expenditure. This seems in tension with the idea that using one's time to grow the number of people who are also impartial altruists counts as investment, such that opportunities to spend one's time in this way do not contribute positively to the influence attaching to a particular moment in history.

Admittedly, there is no strict contradiction here. There are ways of achieving desirable persistent changes in values that need not involve increasing the number of impartial altruists, such as getting people to see that judicial torture should be abolished. But it is natural to wonder why one would wish to draw a line here.

More generally, when MacAskill says that 'investment' refers "to both financial investment, and to using one's time to grow the number of people who are also impartial altruists" (335 n.13), does he intend to say that it refers to just these two things? Consider, say, spending money, rather than time, to grow the number of people who are impartial altruists or using one's time to increase the influence of impartial altruists but not their number. Are these also examples of investment, given that using one's time to grow the number of people who are impartial altruists is counted as such? Intuitively, these things should go together.

If the concept of investment is enlarged in this way, the natural next question is what it includes in its full generality or what principle can be used to determine as much. I believe no clear answer to this question is suggested in MacAskill's discussion. Furthermore, one natural answer is foreclosed to us. The natural answer I have in mind is as follows. It says that an investment involves using one's resources in some way for the sake of achieving a desirable

outcome at some suitably distant future time, rather than some immediate payoff. Why do I say that this definition is foreclosed to us? I say that because MacAskill's discussion presupposes *longtermism*. It presupposes something like the view that "far future effects are the most important determinants of the value of our options." (Greaves and MacAskill 2021: 3) Consider that when MacAskill argues that we are more influential than someone living in Europe in the Early Modern period, he writes: "the opportunities available to this person in 1600 were in general less high-leverage than the opportunities available to us today. In particular, they would have had few opportunities to shape the long-run future" (346-7). By the definition just proposed, using our resources to achieve desirable effects in the far future counts as an investment. Therefore, opportunities to allocate resources in this way do not count positively toward the influence attaching to a given time.

Note, moreover, that by the proposed definition, there is no question for longtermists as to whether we should prioritise investment. If longtermism entails that we should allocate resources so as to achieve desirable effects in the far future, then it immediately entails that we should make investment our top priority. But MacAskill claims that there is only "a *prima facie* presumptive argument" (335) from longtermism to the conclusion that we should be investing our resources, which may be overturned by evidence that the current era is uniquely influential.

The upshot is that I don't feel confident that I know how to interpret the distinction between direct expenditure and investment that MacAskill has in mind. Nor is this the only problem that attaches to his use of this distinction.

By MacAskill's definition, influence is tied to how much expected good one can do with the direct expenditure *rather than* investment of a unit of resources. It's not clear whether this should be taken to mean that t_1 is more influential than t_2 just in case one can do more

good in expectation via direct expenditure at t_1 than at t_2 or whether it should instead be taken to mean that the difference in the expected good of allocating resources to direct expenditure as opposed to investment is greater at t_1 than at t_2. My impression is that the reasoning used to support MacAskill's inductive argument relies on the first of these definitions, because MacAskill does not appear to compare the difference in expected good achievable by direct expenditure and investment for the historical times he considers. But this seems to be the wrong approach, insofar as the decision we face is whether to spend or invest now. Suppose that the good that can be achieved in expectation by investment also rises over time and has consistently remained exactly equal to the good that can be achieved in expectation through direct expenditure. The conclusion that is inductively supported is that the expected value of direct expenditure is no higher than the expected value of investment today. We would be misled if we inferred that since there exists an upward trend in the expected good achievable through direct expenditure, the best thing for us is to save for tomorrow.

There are two additional sources of unclarity worth highlighting. As you may recall, a central part of the inductive argument rests on the claim that past people were less well-placed to have a positive impact because of the significant gaps in their empirical knowledge. Viewed in a certain perspective, lack of knowledge need be no impediment to how much good a person can do in expectation. Suppose we know that if Anne had only taken the long road home, she would have seen Bob drowning in the lake and would almost surely have saved him. Since she took the shorter route, she never learned of his peril. We might naturally say that Anne could probably have saved Bob, she just didn't know it.

The issue here relates to the role played by the concept of *expected value* in MacAskill's definition. Assume that we are Bayesians and so understand probabilities as measures of

coherent degrees of belief. Since Bayesian probabilities are subjective, we need to say whose probabilities are relevant to determining how much good they could have done in expectation. With respect to Ann, the measure of how much good she could have done in expectation on her drive home is high relative to the probability function warranted by our evidence, but not relative to the probability function warranted by Ann's.

Consider, then, some past era. In determining the expected value of the expenditure of a unit of resources during that era, should we use the beliefs of people at the time? Or should we use our own beliefs? For MacAskill's inductive argument to work, it seems we have to use the beliefs of past people. The subjective probabilities of the reader incorporate knowledge, based on hindsight and a scientific education, of things that past people could not have known and in light of which their actions are a great deal more influential than may have been apparent. However, when addressing this issue explicitly, MacAskill asserts that "the probability distribution that goes into the idea of 'expected value' in the definition of influentialness is our own." (337)

It may be thought that we should just stick to using the subjective probabilities of past people in determining the expected value of the best acts available to them. However, this yields strange results in some cases. Consider the Aztec priests who conducted the last New Fire ceremony in 1507, kindling a fire on the chest of a sacrificial victim to mark the end of a 52-year period in their calendar. The priests believed that "if a fire could not be drawn, then [the sun] would be destroyed forever; all would be ended; there would evermore be night." (quoted in Smith 2012: 237) These belief might have been warranted by evidence they were rational in trusting, based on the testimony of the recognized loci of epistemic authority in their society. It seems strange to count this as one of the most important points in human

history because it was so according to a cosmology we know to be false, even if it was one that people at the time might have rationally affirmed.[4]

Another key component of MacAskill's inductive argument appeals to the impoverished moral knowledge of past people. Here too, I think we find an apparent contradiction in how the concept of influence is handled. Recall that MacAskill defines a time as influential in proportion to "*how much expected good one can do with the direct expenditure* (*rather than investment*) *of a unit of resources*" (335) Note that what is key here is how much expected good one *can* do. MacAskill himself highlights this point, stating that it is "worth emphasizing" that "one's influentialness is given by how much expected good one *can* do at a time. It is not given by how much (expected) good one *actually* does." (337) This does not seem to fit with the claim that the most important factor that explains the greater influence of later individuals vis-à-vis earlier individuals is moral progress. On its face, mistaken moral beliefs should not reduce the expected good that *could* have been achieved in expectation at the time. It only makes it less likely that the morally best options would be chosen. MacAskill must either revise his definition of what it means for a time to be influential or give up what he claims to be the most important consideration supporting the conclusion that our influence has been increasing over time.

4. Conclusion

While MacAskill has made significant strides in bringing academic rigour to bear on the claim that we are living at the 'hinge of history,' he fails to make a convincing case against this

[4] This example is based on a case suggested to me by Christian Tarsney.

hypothesis. His first argument relies on interpreting the claim that we live at the 'hinge of history' in a way that does not match what those who assert this claim have had in mind. His conception of what it means for a time to be influential is too unclear to support his inductive argument. Of course, the authors he criticizes are even less clear on what conception of importance they have in mind. When it comes to whether we are living at the 'hinge of history,' the first thing we need is greater conceptual clarity. At present, we hardly understand the question, much less how to answer it.

Bibliography

Bostrom, Nick (2002) *Anthropic bias: observation selection effects in science and philosophy*. London: Routledge.

Bostrom, Nick (2014) *Superintelligence: paths, dangers, strategies*. Oxford: Oxford University Press.

Curtin, Ciara (2007) Fact or fiction?: Living people outnumber the dead. *Scientific American* <https://www.scientificamerican.com/article/fact-or-fiction-living-outnumber-dead/>

Greaves, Hilary and MacAskill, Will (2021) The case for strong longtermism. *Global Priorities Institute Working Paper No. 5-2021.* < https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf>

MacAskill, Will (2019) Are we living at the most influential time in history? *Effective Altruism Forum* <https://forum.effectivealtruism.org/posts/XXLf6FmWujkxna3E6/are-we-living-at-the-most-influential-time-in-history-1>

MacAskill, Will (2022) Are we living at the hinge of history? In McMahan, Campbell, Goodrich, and Ramakrishnan, eds. *Ethics and existence, the legacy of Derek Parfit*, 331-57. Oxford: Oxford University Press.

Ord, Toby (2020) *The precipice: existential risk and the future of humanity*. London: Bloomsbury.

Parfit, Derek (2011) *On what matters, vol. 2*. Oxford: Oxford University Press.

Sagan, Carl (1994) *Pale blue dot: a vision of the human future in space.* New York, NY: Random House.

Smith, Michael E. (2012) *The Aztecs, 3rd ed.* Oxford: Wiley-Blackwell.