The hinge of history hypothesis: Reply to MacAskill

Andreas Mogensen (Global Priorities Institute)

Global Priorities Institute | August 2022

GPI Working Paper No. 9 - 2022

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 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 vanishingly small, 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 trend throughout history is for later generations to be more influential.

In my view, neither of these arguments should convince us. I criticize the first argument in section 2, showing that it relies on formulating 3H in a way that does not conform to how this hypothesis is traditionally understood. I criticize the inductive argument in section 3, noting that MacAskill's definition of what it means for a time to be influential leaves us with important unresolved ambiguities, and that the argument appears to rely on resolving at least one of these ambiguities in a way that seems misguided given MacAskill's practical concerns.

2. MacAskill's Priors Argument

In MacAskill's preferred formulation, 3H states that you and I 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 who may 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. Thus, according to Parfit (2011: 616), 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, an apparent mismatch between MacAskill's formulation of 3H and Parfit's conception. For MacAskill, the claim that we live at the 'hinge of history' builds in the assumption that humanity will continue for long enough to establish a significant presence throughout the galaxy. For Parfit, 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,' zero urgency 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 extinction is especially urgent.

This suggests that MacAskill's formulation distorts the issue - or at least changes the subject. But perhaps this can easily be 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 that its *expectation* is astronomical in size. Since the latter claim can be accepted by someone who assigns significant probability to near-term extinction, the mismatch between MacAskill's formulation and Parfit's conception of the 'hinge of history' would dissolve. However, MacAskill's first argument against 3H cannot withstand this change in formulation.

This argument 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

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

Thus, if I update by conditionalization, learning that m of n observers in my reference class exhibit F sets my credence that I exhibit F to m/n. For example, learning that 4 of the 20 guests at the party are secretly admired by the host should make me 20% confident that I am so admired. If F is a superlative property, m/n necessarily equals $1/n$ (assuming ties are impossible). Thus, my prior probability that I am the guest most admired by the host should be 5%.

MacAskill argues that SSA entails that the prior probability that you and I 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)

This line of argument assumes MacAskill's chosen formulation of 3H, since it assumes that there will be trillions of people to come. Suppose that we replace the claim that we *will* be succeeded by trillions of people with the claim that the *expectation* of the future population size is astronomical. 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 extremely low. We might suppose that if N is a random variable that denotes the future population size in millions, 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.[1]

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

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

Then, by the Law of Total Probability,

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

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

[1] Something like this line of reasoning is arguably suggested by MacAskill (2019) in an earlier draft: "we could use, say, 1 trillion years' time as an early estimate for the 'end of time' (due to the last naturally occurring star formation), and a 0.01% chance of civilisation surviving that long. Then, as a lower bound, there are an expected 1 million centuries to come, and the natural prior on the claim that we're in the most influential century ever is 1 in 1 million."

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

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 $\mathbb{E}(N)$. This is not too unlike the situation in which we find ourselves if we live at the Parfitian 'hinge of history.' If there is a 10% chance that the current generation of 7.7 billion people is the last, then the value of $\mathbb{E}(1/N)$ is *at least* thirteen orders of magnitude greater than the figure of one in a million trillion suggested by MacAskill for the prior probability that we are among the million most influential people yet to come.

It may be thought that MacAskill's argument can be strengthened if we also look to the past. Conservatively, there have existed at least 100 billion people (Curtin 2007). By SSA, the prior probability that we are among the million most influential people to have existed given that there have existed at least 100 billion people is *at most* 1 in 100,000.

The claim that we live at the 'hinge of history' does typically 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 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 little to 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 provides 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 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 himself 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

According to MacAskill's 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 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, this 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, MacAskill says little to justify thinking that people at earlier times were, in general, more influential than people at still earlier times. His inductive 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 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 appears to rely on resolving one of these ambiguities in a way that is wrong-headed.

Let me begin with two examples of apparent contradictions in MacAskill's deployment of the influence concept, for which a plausible resolution may nonetheless be available. These concern the beliefs that are relevant in assessing the influence of past people.

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 significant gaps in their empirical knowledge. However, viewed in a certain perspective, lack of knowledge need be no impediment to how much good a person can do in expectation, but only to their knowledge of how much good they can do in expectation.

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 someone could have done in expectation. In determining the expected value of the expenditure of a unit of resources during some past era, should we use the beliefs of people at the time or our own beliefs? Given his emphasis on the impoverished empirical knowledge of past people, for MacAskill's inductive argument to work, it is natural to think 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)

Another key component of MacAskill's inductive argument appeals to the impoverished moral knowledge of past people. However, MacAskill defines a time as influential in proportion to "*how much expected good one can do*" (335). MacAskill himself emphasizes 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 is moral progress. On its face, mistaken moral beliefs should not reduce the expected good that *could* have been achieved in expectation at some past time. It only makes it less likely that the morally best options would be chosen.

Here is a possible solution to both problems highlighted above.[3] We measure how influential some past person was in terms of how much good they would have done in expectation had they tried to do the most good – as opposed to how much good they *could* have done or *actually* did – and use our own beliefs in assessing the expected value of the act they would have chosen, while using their own empirical and moral beliefs to determine what option they would have chosen in pursuit of the beneficent aim that we counterfactually attribute to them. This allows us to go between the horns of the dilemma. It seems to capture MacAskill's usage and neatly rationalizes those parts of his inductive argument highlighted above.

One problem with the suggested solution is that it does not incorporate the emphasis on *direct expenditure* built into MacAskill's definition of influence. Had they wanted to achieve the most good, past people might at various times have favoured investment, whereas our interest is in historical patterns that attach to the good achievable through direct expenditure

[3] I'm grateful to an anonymous referee and Ben Eggleston for this suggestion.

rather than investment. By way of an easy fix, we might consider how much good past people would have achieved had they aimed to do the most good *specifically as a result of direct expenditure*. Depending on how we conceive of direct expenditure, this may involve counterfactually attributing a somewhat abstruse or unnatural goal to past people, since we seem required to imagine them as caring specifically about direct expenditure as a vehicle for beneficence. But that is not obviously damning. The deepest problems with MacAskill's argument arise, I claim, from the significant unclarities that attach to the distinction between direct expenditure and investment in the first place.

This distinction is also 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 to achieve persistent changes in values counts as direct expenditure. However, 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), it is natural to wonder whether 'investment' is meant to refer to just these two things and nothing else. 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 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 determinant 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, and which I have argued is crucial to the success or failure of his argument. 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 he thereby means 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 – ignoring investment entirely - or instead that the difference in the expected good of allocating resources to direct expenditure rather than investment is greater at t_1 than at t_2. The opportunities of people in the Early Modern period to beneficially shape the long-run future through direct expenditure may well have been meagre compared to ours today, but the same is probably true of their ability to achieve morally good outcomes through investment. Thus, they might in principle be less influential than we are today on the first definition, but not the second.

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, and MacAskill (2022: 332) is clear that he is "choosing to define [his] terms so that they are action-relevant, bearing on the question of whether to 'give now or give later'." Suppose, as seems plausible, that the good that can be achieved in expectation by investment also rises over time. Suppose in fact that it 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 evidence favours saving for tomorrow.

4. Conclusion

While MacAskill has made significant strides in bringing academic rigour to bear on the hypothesis that we are living at the 'hinge of history,' I claim that he fails to make a convincing case against this 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 in mind. His inductive argument ought not to convince us in large part because it remains unclear how to understand the crucial distinction between direct expenditure and investment in light of which a time's importance is defined. 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.

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.