Doomsday rings twice

Andreas Mogensen

Global Priorities Institute | September 2019

GPI Working Paper No. 1-2019

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Doomsday rings twice

This paper considers the argument according to which, because we should regard it as *a priori* very unlikely that we are among the most important people who will ever exist, we should increase our confidence that the human species will not persist beyond the current historical era, which seems to represent a crucial juncture in human history and perhaps even the history of life on earth. The argument is a descendant of the Carter-Leslie Doomsday Argument, but I show that it does not inherit the crucial flaw in its immediate ancestor. Nonetheless, we are not forced to follow the argument where it leads if we instead significantly decrease our confidence that we can affect the long run future of humanity.

1.

Out of everyone who will ever live, how important are you? You might not know how to go about answering this question. Exactly how important were people living in the past, on average, keeping in mind the power they will have wielded over the present age? What resources and challenges will people in the future face? How many future people will there be? These are tough questions.

Even if there isn't much evidence by which to constrain your beliefs, that doesn't give you license to think what you like. It seems sensible to regard it as *a priori* very unlikely that out of all the billions or even trillions of people who might exist over the course of human history, you are among the very most important. Surely only a narcissist would believe something like that about themselves without strong evidence.

The foregoing reasoning hopefully strikes you as intuitively correct. In this paper, I'll present an apparently reasonable argument showing that the presumption that you are *a priori* very unlikely to be among the very most important people who will ever live should increase your confidence that the human species will be wiped out within the next few hundred years, so long as we are allowed to

make a few plausible empirical assumptions about our current historical epoch. I call this argument *Doomsday Redux* (DR).[1]

An informal presentation of the argument is given in section 2. In section 3, I highlight the commonalities between the argument and the infamous *Carter-Leslie Doomsday Argument* (DA) (Carter and McCrea 1983; Leslie 1998; compare Gott 1993, 1994). This allows me to offer a more formal presentation of DR's key steps. Section 4 argues that DR doesn't inherit the key flaw in DA; I argue that DR should be taken as a valid anthropic argument even though DA can be rejected. Section 5 addresses two concerns that might be raised about the argument in section 4. Section 6 considers whether we could reasonably revise our beliefs about our place in human history so as to avoid the gloomy implications of DR. I argue that this could be achieved by significantly decreasing our confidence in our ability to affect the long-run future of humanity.

2.

I said in the previous section that DR rests on certain plausible empirical assumptions about our current historical epoch. As the first step in developing the argument, I'll now spell out these assumptions.

The assumptions I have in mind are those that inform Carl Sagan's (1994) view that the current era represents "the time of perils." (306) According to Sagan, ours is an extraordinary moment in the history of life on earth. This is the first time on Earth that a species has become capable of destroying itself. This is also the first time in Earth's history that a species has acquired the ability to decisively secure its future against global catastrophic risks by spreading to other planets. This view is not Sagan's alone. John Leslie (1998) also characterizes "the next one and a half centuries" as "a period of grave danger," (203) adding that surviving through this period would mean "there will be an excellent chance that the human race will survive for very many further millennia" (203). Derek Parfit

[1] The core intuition driving the argument is not original to me. It was suggested to me by [redacted] and has been discussed informally among [redacted]. Of course, any mistakes in the presentation and discussion of the argument in this paper are mine alone.

(2011) also holds this view, insisting that we live "during the hinge of history." (616) Echoing Leslie and Sagan, Parfit suggests that if we "act wisely in the next few centuries, humanity will survive its most dangerous and decisive period. Our descendants could, if necessary, go elsewhere, spreading through this galaxy." (616)

Suppose it's true that you and I live during a period in which existential risk is unusually high and that this period will be followed by a much, much longer period in which the human race is either extinct or else continues to exist with a robustly minute risk of extinction, something that might be achieved if our descendants spread throughout the solar system or the galaxy. Call this the *Hinge of History Hypothesis* (3H). If 3H is correct, then it seems that you and I *are* unusually important. We each have some chance of decisively shaping the long-run future of humanity by raising or lowering the probability that our species successfully navigates 'the time of perils'. We could do this by campaigning for nuclear disarmament or spreading awareness about risks from synthetic biology. Given just how many people could exist in future, even small changes in the risk of existential catastrophe should be assigned enormous significance (Beckstead 2013; Bostrom 2009, 2013; Parfit 1984: 453-4). Based on very conservative estimates about the possible size of the future population, Bostrom (2013) calculates that "the expected value of reducing existential risk by a mere *one millionth of one percentage point* is at least a hundred times the value of a million human lives." (18-19)

Those who lived not too long ago may have had the power to inadvertently influence humanity's long-term prospects by shaping the emergence of destructive technologies further down the line or the institutional structures that would manage their use, but the inscrutability of future history rendered them unable to do so reliably and predictably. Those who live after 'the time of perils' will have much less resting on their shoulders. If 3H is true, nothing within their power will be nearly as important as the current generation's stewardship of posterity. You and I would therefore be extremely high up in the ranking of the most important people who will ever live.

The foregoing line of reasoning assumes, of course, that there *will* be people living after 'the time of perils'. That is obviously far from certain. Suppose instead that we fail in our stewardship of posterity and the human species dies out within the 'time of perils'. Then our relative position in the

ranking of the most important people who will ever live would not be nearly so impressive. If the human race were to carry on for billions or trillions of years, those who lived during the current historical era would represent an extremely small fraction of all the people who will ever exist. With so many less important generations succeeding us, we would probably be among the top 0.01% most important people in history. But if we go extinct soon enough, then our relative importance won't be nearly so extraordinary. Those of us who lived during the 'time of perils' would represent nearly 10% of all people who have ever lived. *A priori*, it seems far more likely that you and I are in the top 10% of the most important people to have existed, than that we are in the top 0.01%. The idea that it is *a priori* very unlikely that you and I are among the most important people who will ever live therefore apparently supports the chilling conclusion that our species won't make it through this dangerous and decisive period.

2.

The previous section gave us an initial, informal presentation of DR. As noted, DR is a descendant of DA. DA also trades on the idea that we occupy a position that any randomly selected human being would be many times more likely to occupy if the future of humanity is very short as opposed to very long. DA focuses on our birth rank, as opposed to our importance. I'll now offer a formal statement of DA, after which I'll present a similarly rigorous exposition of DR, making clear the fundamental similarities between them.

Assume for simplicity that there are only two hypotheses concerning the total size of the human species to which you assign any credence. H_1 states that there will exist 100 billion human beings. H_2 states that there will exist 100 trillion human beings. We may understand H_1 as corresponding to the view that we will not make it through the 'time of perils,' and H_2 as corresponding to the view that we will. We will also assume the so-called *Self-Sampling Assumption* (SSA) (Bostrom 2002). SSA requires you to reason as if you were a randomly selected sample from the observers within your reference class. Let n be the number of observers in your reference class, of which m instantiate F. Denote yourself by the indexical expression i, and let $\Pr(\cdot)$ denote your rational credence function. Then, SSA entails $\Pr(Fi \,|(m/n) = k) = k$.

Consider then your *birth rank*, i.e., the number of all human beings born prior to you, plus one. Your birth rank is approximately 60 billion. Let's assume for simplicity that it has exactly this value. Conditional on H_1, a randomly selected human being is far more likely to have this birth rank than if we consider the same conditional probability relative to H_2. Letting F be your birth rank, SSA entails that $\Pr(Fi \,|H_2)/\Pr(Fi|H_1) = 0.001$.

Suppose that, in light of the available non-indexical evidence, you are antecedently more confident in H_2. In other words, you are very confident that humanity will have a long and glorious future, taking account of the available object-level evidence relating to the extinction risks we are likely to face and setting aside any applications of anthropic reasoning. For the sake of argument, let your prior confidence so described be such that $\Pr(H_2) / P(H_1) = 19$. Then, by application of Bayes' rule, we find that $\Pr(H_2 \,|Fi\,)/\Pr(H_1|Fi\,) = 0.019$. Conditionalizing on your birth rank in accordance with the prior required by SSA will change your confidence so that whereas you were previously only 5% confident in H_1, you are now more than 98% confident in this hypothesis. The self-locating evidence represented by your birth order apparently requires a significant probability shift in favour of near-term extinction.

The basic structure of DR is the same, except that the argument is driven not by consideration of your birth-rank, but by consideration of your importance.

What exactly do I mean by importance? Here is a proposed precisification of this notion. Consider all the different options available to an agent. For each action, we consider the possible consequences of performing that action and determine the absolute value of the improvement or decrement in the total value of human history that would result if that consequence were to result from performance of that action, as opposed to if the agent were to exercise her agency so as to make no difference to the run of events. We weight these value-differences in terms of the probability on the agent's evidence that that consequence would in fact result from performance of the action. We sum these weighted value-differences to obtain the expected value-difference associated with any given action. One agent is more important than another if the action available to the first agent whose expected value-difference is greatest has an expected-value difference higher than that of the action available to the second agent whose expected value-difference is greatest.

Importance, so understood, is not tied to success. If you can significantly affect the probability of human extinction, you are very important. This holds true even if humanity goes extinct within your lifetime because you refuse to make the effort. Furthermore, you will count as important even if you make the best of your powers but still humanity would have survived even had you not acted. What matters is that you were able to sufficiently alter the probability of a near-term doomsday.

Since we live during the 'time of perils,' we may conclude that you and I are extremely important as human beings go. Denote the property of being at least as important as we are as *G*. Assume for simplicity that 10 billion people will live during 'the time of perils' and they will all instantiate *G*. Assume furthermore that only these people instantiate *G*, regardless of whether H_1 or H_2 is true. In other words, we assume that the people who will live after us are less important than we are. By virtue of living after 'the time of perils,' nothing within their power will be nearly as important as the current generation's stewardship of posterity.

Given these assumptions, SSA entails $\Pr(Gi\,|H_2)/\Pr(Gi|H_1) = 0.001$, because a randomly selected person is 1,000 times more likely to instantiate *G* conditional on H_1 than on H_2. Therefore, by reasoning parallel to DA, even if you are antecedently very confident that we will survive the 'time of perils,' conditioning on the fact of your historical importance will shift your posterior credence significantly toward the view that we will not endure the next few centuries. Like before, suppose your prior confidence is such that $\Pr(H_2) / P(H_1) = 19$. In other words, you antecedently assign 5% credence to H_1 and 95% credence to H_2. Given entails $\Pr(Gi\,|H_2)/\Pr(Gi|H_1) = 0.001$, by application of Bayes' rule, we find that $\Pr(H_2\,|Gi\,)/\Pr(H_1|Gi\,) = 0.019$. In other words, updating on the fact of your importance via SSA means that your posterior credence in H_2 is approximately 1.9%, whereas you attach approximately 98.1% confidence to H_1.

4.

The similarity between DA and DR represents a liability for DR. Very few people are persuaded upon consideration of DA that taking account of anthropic selection effects should lead us to significantly increase our confidence in near-term extinction. Given their similarity, we might expect that DR and DA stand or fall as one. If so, this would be bad news for DR. However, this section will argue that

things are not so. We should reject DA, but the reason why we should reject DA is not a reason to reject DR.

Objections to DA are legion (see Richmond 2006 for an overview). Many can be easily dismissed (Bostrom 2002: 109-126). By my lights, there are two key lines of objection to DA that have emerged in the literature. My discussion will focus on just one of these. Before we get there, I will briefly explain why I set aside the other.

The line of objection that I set aside is due to Bostrom (2002). He maintains that DA is undermined by its fragility under varying assumptions about the reference class. SSA asks you to calibrate your credence that you exhibit some property against the relative frequency with which that property occurs. But any token instance can be subsumed under many different reference classes in which the relative frequency of the property differs, and there is often no principled means of deciding between the available options. According to Bostrom, reasoning that relies on SSA is cogent when its conclusions depend on very weak assumptions about the reference class. Bostrom (2002: 104-7) argues that DA depends on very strong assumptions about the reference class. He therefore concludes that it lacks the robustness characteristic of scientific objectivity.

Very briefly, here is why I am not convinced by this line of reasoning. Any plausible theory of inductive logic must allow that agents can rationally update their non-indexical beliefs in light of facts about frequencies. Taking account of frequencies requires adopting suitable reference classes by which to classify token events, rejecting others as inappropriate. Justifying the choice of any particular reference class is often non-trivial, as the grue paradox makes clear (Goodman 1983). I think the inescapability of the reference class problem in inductive logic means that Bostrom's objection to DA is not a very strong one. Proponents of DA have too many eligible partners in crime for the charge to stick (compare Hájek 2007). I obviously don't intend these cursory remarks as a refutation of Bostrom, but merely as a reasonable explanation for why I'll say no more about this line of objection throughout the rest of this paper.

In my view, DA actually fails because your birth rank is screened off as evidence for H_1 by prior knowledge that you occupy a spatiotemporal location occupied by exactly one observer regardless of which out of H_1 or H_2 is correct. I believe this conclusion emerges naturally from what I

regard as the second key line of objection to DA to have developed in the literature: namely, that the probability shift in favour of near-term extinction obtained by conditionalizing the prior required by SSA on our birth-rank ends up being cancelled out by adjusting our priors in accordance with a different principle of anthropic reasoning, known as the *Self-Indication Assumption* (SIA) (Dieks 1992, Bartha and Hitchcock 1999, Olum, 2002, Monton 2003).

Before we proceed, there is just one final bit of stage-setting to note. My description of the appeal to SIA as a 'line of objection to DA' may be questioned. Philosophers who appeal to SIA in responding to DA typically do not deny that application of SSA to our birth-rank should lead us to update in favour of near-term extinction. They merely insist that this shift is cancelled out by application of SIA. This constitutes an objection to DA only insofar as we understand DA as purporting to show that we ought to be (significantly) more confident that the human species will not survive 'the time of perils' once we take into account all relevant principles of anthropic reasoning. If we adopt a weaker reading, on which DA merely purports to show that our birth-rank provides evidence against H_2 when we take into account the application of some valid principle of anthropic reasoning, then appeal to SIA of itself provides no objection to DA. However, the position I develop purports to show that DA fails even on this weaker reading.[2] And my claim is that this insight emerges naturally from attempting, by appeal to SIA, to show that DA fails on the stronger reading. By contrast, I will argue that DR may be used to successfully support analogues of both the weaker *and* the stronger conclusions just noted.

Let's now proceed to clarifying exactly what SIA requires of your credences. Roughly speaking, SIA asks you to assign greater credence to worlds with more observers. If you and I had reason to think that we were the only conscious minds that would ever emerge and behold the universe in all its splendour, we would probably think ourselves very lucky. SIA may be thought of as

[2] More exactly, I will argue that DA fails (even on this weaker reading) provided that we assume (very plausibly) that relevant knowledge about your spatiotemporal location is obtained prior to knowledge of your birth rank, without which the screening off effect I've mentioned obviously wouldn't occur. For further discussion of this assumption, see the second half of section 5.

formalizing that intuition. For any hypotheses, H_i, and H_j, if m is the number of observers in one's reference class existing in H_i and n is the number of observers existing in one's reference class in H_j, then SIA requires you to set your priors so that $\Pr(H_i) / \Pr(H_j) = m/n$.

Considered individually, SSA and SIA both seem to allow you to do too much with too little, using apparently trivial evidence or no evidence at all to significantly raise or lower the probability of more populous worlds. But taken together, they cancel out, leaving you none the wiser. In other words, if we alter our priors so as to accord with SIA and then update on the evidence of our birth rank in accordance with the prior required by SSA, we find that the overall effect is nil and our confidence in H_1 vis-à-vis H_2 is exactly as it was before the application of these anthropic principles. In this way, we seem to satisfy our intuitive expectation that you can't really get so much from so little. Notably, no assumptions about the reference class are required in order to achieve this cancelling effect, so long as the same reference class is used in applying SIA and SSA.

Unfortunately, the story about DA can't be quite so simple as this. A well-known concern about SIA is that the conditions of its applicability are not constrained to those in which you know your birth rank and can thereby cancel out prior confidence in more populous worlds by application of SSA. In those conditions, SIA appears to allow you to be absurdly confident in hypotheses favouring the existence of more populous worlds. By way of example, consider the following case due to Bostrom (2002):

The Presumptuous Philosopher

In the near future, physicists have only two remaining candidates for the Theory of Everything: T_1 and T_2. T_1 entails that there are a trillion trillion observers in the cosmos. T_2 entails that there are a trillion trillion trillion observers. The physical evidence currently does not favour either theory, but an inexpensive experiment just about to be run is expected to settle the matter. A philosopher contacts the physics community and informs them that they needn't bother running the experiment: its outcome can already be known with near certainty, since SIA requires us to treat T_2 as one trillion times more likely than T_1.

It seems clear that something has gone wrong here. However, the error in the philosopher's reasoning can't be one of failing to apply SSA to her knowledge of her birth-rank so as to cancel out her greater prior confidence in T_2 induced by SIA. The philosopher is in no position to know her birth-rank, even approximately. This may suggest that something is wrong with SIA itself.

A partially convincing reply to the *Presumptuous Philosopher* objection was outlined by Monton (2003). Its general point is conveniently reinforced by DR. A shift in favour of a smaller total population estimate induced by conditionalizing the prior required by SSA does *not* require the evidence on which you update to be knowledge of your birth rank. The same shift can be induced if you can identify other properties that you exhibit, of which you know that the same number of observers will exhibit that property regardless of which of two different hypotheses about the total population is correct. An example is having the property *G*, i.e., having a level of importance at least as great as the high level of importance that we have by virtue of living during the 'time of perils.' Monton infers that it *is* possible to run a DA-style argument even in the *Presumptuous Philosopher* case. The philosopher can run such an argument by conditionalizing on the number of observers existing within some appropriate space-time region within which she finds herself. We assume she knows how many observers are there. Furthermore, she knows that if they are to count as potential theories that correctly describe the universe and everything within it, both T_1 and T_2 must agree on how many observers exist in that region.

Talk of space-time regions is arguably too hifalutin in this context. Really, all the philosopher need know is the trivial indexical proposition *I am here now*, provided that *here now* rigidly designates a point she knows is occupied by herself and no one else. So long as the philosopher knows of her indexically characterised spatiotemporal location that it is uniquely occupied - and uniquely occupied whether T_1 or T_2 is correct - she is able to reason in the manner described by Monton. Given SSA, the prior probability that the philosopher should have assigned to occupying the spatiotemporal location that she knows herself to occupy under the rigidly designating mode of presentation *here now*

is one trillion times greater on T_1 than T_2.[3] The philosopher therefore does have evidence of the kind she would need in order to apply SSA and cancel out the greater prior probability of T_2 induced by SIA. Moreover, it seems plausible that there are no realistic conditions under which an observer is unable to conditionalize on indexical knowledge of her uniquely occupied spatiotemporal location so as to cancel out the greater prior confidence in more populous worlds induced by SIA.

However, this reply to the *Presumptuous Philosopher* objection may seem to let DA in via the backdoor. That is why I say that it is only partially convincing. Return to the case in which we are uncertain between H_1 and H_2. By the reasoning of the foregoing paragraph, conditionalizing on *I am here now* via SSA favours H_1, exactly cancelling out my greater prior confidence in H_2, as required by SIA. What happens, then, when I learn my birth rank and apply SSA to that knowledge? The greater prior probability for more populous worlds induced by SIA has already been cancelled out. Therefore, it seems that there is nothing to counterbalance the confidence I ought to gain in the hypothesis of near-term extinction by conditionalizing on my birth rank via SSA. We seem to be right back where we started.

The solution to this problem emerges from the principle that where one body of evidence, E_2, is entailed by another, E_1, prior knowledge of E_1 screens off E_2. This principle is equivalent to the following theorem of the probability calculus (assuming all relevant conditional probabilities are well-defined with non-zero denominators): $[E_1 \vDash E_2] \rightarrow [\Pr(H \mid E_1 \wedge E_2) = \Pr(H \mid E_1)]$. By way of illustration, suppose we know that Madhav reports that a mugger wore a blue jumper. This entails that Madhav reports either that the mugger wore a blue jumper or that the mugger was a green elephant. If I knew only this disjunctive fact, I would increase my confidence that the mugger wore a blue jumper. However, knowledge of the disjunction doesn't increase the probability of that hypothesis when I already know the content of Madhav's report, because the disjunction is entailed by my prior evidence.

[3] Obviously, she was certain to occupy some spatiotemporal location that she would be able to designate as *here now*. But she need not have ended up at the location to which her actual use of *here now* refers. She might conceivably have ended up somewhere else.

How does this help us with the problem noted two paragraphs ago? Note, first of all, that if your spatiotemporal location is to serve as evidence for H_1, you must occupy a spatiotemporal location that is occupied if H_1 is true. In other words, you must occupy a point in time prior to that point in time at which people who exist only if H_2 is true come into being. This entails that you have a birth rank less than that of any person who exists only if H_2 is true (Bradley 2005).[4]

Note also that this is the weakest fact encoded by your birth rank in light of which the conditional probability of H_1 is many times greater than that of H_2, given SSA. In other words, the probability that a randomly selected person will have a birth rank less than that of any person who exists only if H_2 is true is many times greater conditional on H_1 than on H_2. Denote the property of having such a birth rank as *K*. Since everyone who exists if H_1 is true exhibits *K* but only 0.1% of those who exist if H_2 is true exhibit *K*, we have $\Pr(Ki\,|H_2)/\Pr(Ki|H_1) = 0.001$.

Note, finally, that knowledge of your *exact* birth rank does not alter the likelihood ratio. As we recall, where *F* denotes the property of having birth-rank 60 billion, $\Pr(Fi\,|H_2)/\Pr(Fi|H_1) = 0.001$. Because the likelihood ratio is unchanged, we infer that $\Pr(H_1|Fi) = \Pr(H_1|Ki)$. Furthermore, since *Fi* entails *Ki*, $\Pr(H_1|Fi \wedge Ki) = \Pr(H_1|Fi)$. Hence, $\Pr(H_1|Fi \wedge Ki) = \Pr(H_1|Ki)$. Gaining knowledge of your particular birth rank provides no additional evidence in favour of H_1 once you already know that you exhibit *K*.

[4] Enlisting Bradley (2005) at this point may surprise some readers, as Bradley *defends* DA. His concern is that if a DA-style argument can be run in a situation where you don't know your birth rank and all you have to go on is the fact of your uniquely occupied spatiotemporal location, then DA is subject to an even more powerful version of the concern that it allows you to do too much with too little. He does not contest Monton's observation that you can induce a DA-style shift by application of SSA to any property that you know you exhibit and that the same number of observers will exhibit whether H_1 or H_2 is true. Instead, Bradley argues that whenever you know that your uniquely occupied spatiotemporal location is uniquely occupied whether H_1 or H_2 is true, you are also in a position to know the relevant facts about your birth rank. Hence, the knowledge-base of the traditional DA is still available to you in those cases. While fully agreeing with this observation, I think of it as a Pyrrhic victory. While Bradley's observation may help to address the particular concern that motivated his paper, I believe it ultimately undermines DA, for the reasons I describe in this section.

Suppose, then, that you have already updated in favour of H_1 by conditionalizing on the knowledge that your indexically characterized spatiotemporal location is occupied by exactly one person. As we've seen, the evidence on which you have updated entails *Ki,* and hence learning *Ki* does not further increase the probability of H_1. Nor, we've seen, does learning your exact birth rank increase the probability of H_1 by any more than it was increased by learning *Ki*: i.e., not at all. Therefore, learning your birth rank does not further increase the probability of H_1. And that is why DA fails. *QED.*

I think it's useful to note that this screening off effect doesn't always happen when we consider whether knowledge of your birth rank provides evidence for smaller populations. Consider the *Presumptuous Philosopher* case yet again. The philosopher knows *I am here now,* but she does *not* know that her birth rank is less than that of any person who exists given T_2 but not given T_1. She doesn't know her birth rank. Therefore, learning her birth rank could alter her credence between the two theories, redistributing probability mass from T_2 to T_1. That seems exactly right. For example, suppose she learns that her birth rank is just what it would be if only the observers in T_1 existed. Intuitively, she should then update in favour of T_1. Here is a case where there seems to be nothing fishy about updating in favour of less populous worlds given knowledge of your birth rank - precisely because there is no assumption that anyone who exists only if the more populous hypothesis is true must have a birth rank greater than your own.

Now that we know what's wrong with DA, let's consider whether a similar objection applies to DR. The same objection would apply to DR only if knowledge that your indexically characterized spatiotemporal location is uniquely occupied allowed you to infer facts about your importance: in particular, the fact that you have an importance level greater than that of anyone who exists only if H_2 is true. But knowing that your indexically characterized spatiotemporal location is uniquely occupied clearly does not allow you to infer that you are more important than anyone who will exist only if H_2 is true. Therefore, the fact of your importance is not screened off as evidence for the hypothesis of near-term extinction in the same way as your birth rank.

Note, moreover, that the probability shift in favour of near-term extinction induced by application of SSA to this piece of evidence cannot be cancelled out by appeal to SIA without double

counting. That is because the greater prior probability for more populous worlds induced by SIA has already been cancelled out by the application of SSA to my knowledge of my uniquely occupied spatiotemporal location characterized under an indexical mode of presentation.

5.

In this section, I want to quickly address two potential points of confusion regarding the argument in the previous section.[5]

First of all, how can I say that my spatiotemporal location doesn't enable me to infer facts about my importance relative to other human beings? After all, the reason given at the outset for thinking that you and I are unusually important as human beings go is the truth of 3H: i.e., that we live (on Earth!) during 'the time of perils.' Haven't I been inferring something about our relative level of importance by appeal to where we find ourselves in space and time?

In response, I note first of all that 3H does not strictly entail that we have an unusually high level of importance that any randomly selected person is 1,000 times less likely to exhibit given H_2 as opposed to H_1. It's consistent with 3H that there's nothing we can do to affect the probability of a good long-run outcome for humanity. If we're powerless in that way, we wouldn't be especially important. We would be mere spectators to the great drama. The level of importance that we exhibit would then be one that we should expect to be reasonably common among any generations that succeed us, neutralizing DR. (We'll return to this point in section 6.)

More importantly, I don't claim that knowledge of my spatiotemporal location cannot enable me to infer the relevant facts about my level of historical importance. Instead, I claim that knowledge of my *indexically characterized* spatiotemporal location cannot do so. The mode of presentation is key. Knowing *I am here now* is not the same as knowing *I exist on Earth during 'the time of perils'*, even if *here now* rigidly designates a spatiotemporal location on Earth during 'the time of perils'. The reason DA fails is that knowing *I am here now* and that exactly one person is *here now* regardless of

[5] I'm grateful to [redacted] for alerting me to these issues, though I remain unsure that I've successfully captured their concerns.

whether H_1 or H_2 is true screens off knowledge of my birth-rank as evidence for H_1, since it entails that my birth rank is less than that of anyone who exists only if H_2 is true. I claim that the same does not hold for DR. That is compatible with thinking that knowing my uniquely occupied spatiotemporal location under some other, more informative mode of presentation would screen off my knowledge of my importance as evidence for near-term extinction.

Here's a second question about the argument in section 4 that I want to address. Am I saying that the probability shift in favour of near-term extinction induced by conditionalizing on the fact of my importance via SSA cannot be cancelled out by adjusting my priors to accord with SIA, but the shift induced by conditionalizing on my uniquely occupied spatiotemporal location *can* be cancelled out in this way? This may seem puzzling. Why is the one shift cancellable and the other not?

As I see it, two pieces of evidence support the hypothesis of near-term extinction via the application of SSA: my uniquely occupied spatiotemporal location and my importance. The support offered for the hypothesis of near-term extinction by either of these pieces of evidence can in principle be cancelled out by appeal to SIA, but SIA cannot be relied on to cancel out the evidentiary significance of both without double-counting.

I have assumed, quite plausibly, that the fact of my importance is something I discover after having taken note of my indexically characterized spatiotemporal location. If the order of discovery were inverted, things would be different. In that case, the support for near-term extinction offered by the fact of my importance would have merely served to cancel out the greater prior probability for more populous worlds induced by SIA. Subsequently coming to know of my uniquely occupied location under its indexical mode of presentation should then increase my confidence that doom will come soon, without this shift in probabilities being cancellable by adjusting my priors to accord with SIA. But it is hard to imagine any realistic scenario under which the order of discovery would run in that direction.

6.

From this point onward, I'll set aside further comparisons between DR and DA. I will assume that DR represents a valid anthropic inference, and that the probability shift it mandates cannot be cancelled

out by appeal to SIA (at least not in any realistic scenarios). The question I want to consider in this final section is whether we are forced to follow the argument where it leads, or whether there are other changes in our beliefs that we could make such that our historical significance should no longer be viewed as of a kind that a randomly selected person is far less likely to fall under if the human species survives 'the time of perils' than if it goes extinct during this era, thereby undercutting DR.

The first possibility we might consider is that although existential risk is higher now than at any point in history, the risk of extinction will remain roughly as high as it is now for as long as creatures like us continue to exist. There will be no precipitous drop a few centuries hence. A glorious, long-run future will never be guaranteed. Most human beings will live in the shadow of doomsday.

This might allow us to say that our historical position with respect to existential risk isn't one that a randomly selected person would be significantly less likely to occupy if the human species survives 'the time of perils,' but there is actually not all that much difference between this hypothesis and the apocalyptic predictions on which we've focused so far. It's plausible that the risk of existential catastrophe within this century is around 15% (Sandberg and Bostrom 2008). Suppose it is 15% forever. Then forever comes crashing down before too long. There will be a greater than 50% chance of extinction within 450 years, and a greater than 99% chance of extinction within 3,000 years.

But perhaps we needn't think that there will be a constant 15% extinction risk per century going forward in order to count ourselves as ordinary and average in terms of the extinction threat level we face, even assuming that humanity enjoys a long and glorious future as opposed to a sudden demise. The extinction threat level that we currently face could be close to average within our reference class even if the risk of extinction drops to and remains robustly fixed at an extremely low value for a very long time within about two centuries. We just need it to shoot back up again later on. In fact, a rise of this kind seems inevitable. After 10^{14} years, the stars will stop shining, and after 10^{40} years carbon-based life will become physically impossible due to proton decay (Adams 2008). The risk of extinction might be extremely low up to the point at which we approach astrophysical thresholds of this kind, at which point the risk of extinction will increase significantly. If the increase is suitably great and there are sufficiently many people alive at that point in time, then the long-term

average extinction risk experienced by observers in our reference class might not be so different from 15%. In order to count as suitably ordinary within the long history, we therefore needn't reject the hypothesis that we are at the precipice of a very long period in which the risk of existential catastrophe is negligible.

It's one thing, however, for the risk of extinction that we face right now to be about average, and something else for our level of importance to be about average. The astrophysical processes that threaten humanity's survival in the far future may be as inevitable as the march of time. Even if they can be mitigated, the expected number of future lives whose existence may be ensured as a result of overcoming these challenges will almost certainly be smaller by many orders of magnitude than the number of flourishing future lives that we expect to shepherd into being by successfully navigating the present age of technological peril. Thus, if our species does survive for that long, we would still count as highly unusual in terms of our importance - and far more unusual than we would have been if humanity had failed to achieve it cosmic potential.

This can be fixed by making a different revision in our beliefs about our place in history. As was noted previously, we could be less important than we seem to be, in spite of the truth of 3H, if our capacity to affect the long-run future of humanity is even more limited than it already appears. Now, the ability to affect even a minute change in the risk of existential catastrophe would make us very important. But perhaps even minute changes are beyond our power. The complexity and scale of the problems that we face may be too great. Perhaps anything you and I might do to try to tip the odds washes out as it filters through the complex networks and sluggish institutions that structure our world. Maybe we are really as powerless in the face of today's existential risks as a herd of triceratops glimpsing the Chicxulub asteroid hurtling toward the Gulf of Mexico. If we revised our beliefs so as to regard ourselves as almost entirely powerless, then our level of historical importance would plausibly be one that a randomly selected person would not be significantly less likely to exhibit even if our species survives the 'time of perils'. In that case, DR would be undermined.

7.

So, should we follow DR where it leads? Or is the more sensible response to suitably revise our beliefs about our ability to influence the long run future, so as to avoid being forced to significantly increase our confidence that the end is nigh? I don't know, though I lean toward the latter. I must also profess ignorance of the ethical implications of my argument. Upon suitably updating our beliefs, would the expected value of actions aimed at mitigating extinction risk diminish to such an extent that these actions should no longer rank highly on our list of priorities? I'm not sure. There is a lot more to discuss. The most important questions still need to be settled. But if I have succeeded in making these live questions for you, then I'll have done all that I set out to achieve in this paper.

References

Adams, Fred C. (2008) Long-term astrophysical processes. In Bostrom and Cirkovic, eds. *Global catastrophic risks*, 33-47. Oxford: Oxford University Press.

Bartha, Paul and Hitchcock, Christopher (1999) No one knows the date or the hour: an unorthodox application of Rev. Bayes's theorem. *Philosophy of Science* 66, S339-S353.

Beckstead, Nick (2013) On the overwhelming importance of shaping the far future. PhD thesis: Department of Philosophy, Rutgers University.

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

(2003) Astronomical waste: the opportunity cost of delayed technological development. *Utilitas* 15, 308-14.

(2013) Existential risk prevention as a global priority. *Global Policy* 4, 15-31.

Bradley, Darren (2005) No Doomsday Argument without knowledge of birth rank: a defense of Bostrom. *Synthese* 144, 91-100.

Carter, Brandon and McCrea, William H. (1983) The Anthropic Principle and its implications for biological evolution. *Philosophical Transactions of the Royal Society A* 310, 347-363.

Dieks, Dennis (1992) Doomsday - or: the danger of statistics. *Philosophical Quarterly* 42, 78-84.

Goodman, Nelson (1983) *Fact, fiction, and forecast,* 4td ed. Cambridge, MA: Harvard University Press.

Gott, J. Richard (1993) Implications of the Copernican principle for our future prospects. *Nature* 363, 315-9.

(1994) Future prospects discussed. *Nature* 368, 108.

Hájek, Alan (2007) The reference class problem is your problem too. *Synthese* 156, 563-585.

Leslie, John (1998) *The end of the world: the science and ethics of human extinction.* London: Routledge.

Monton, Bradley (2003) The Doomsday Argument without knowledge of birth rank. *Philosophical Quarterly* 53, 79-82.

Olum, Ken (2002) The Doomsday Argument and the number of possible observers. *The Philosophical Quarterly* 52, 164-184.

Parfit, Derek (1984) *Reasons and persons*. Oxford: Oxford Unviersity Press.

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

Sandberg, Anders and Bostrom, Nick (2008) Sandberg, Anders and Bostrom, Nick (2008) Global catastrophic risks survey. <http://www.global-catastrophic-risks.com/docs/2008-1.pdf>