A Meta-Doomsday Argument: Uncertainty About the Validity of the Probabilistic Prediction of the End of the World

Draft, version 0.91
– main conclusions likely will stand, but some errors are possible

Alexey Turchin
Digital Immortality Now
Foundation Science for Life Extension
alexeiturchin@gmail.com

Abstract: Four main forms of Doomsday Argument (DA) exist—Gott's DA, Carter's DA, Grace's DA and Universal DA. All four forms use different probabilistic logic to predict that the end of the human civilization will happen unexpectedly soon based on the short duration of observed previous human history. There are hundreds of publications about the validity of the DA. Most of the attempts to disprove the DA have some weak points. As a result, we are uncertain about the validity of DA proofs and rebuttals. In this article, a meta-DA is introduced, which uses the idea of logical uncertainty over the DAs validity estimated based on a virtual prediction market of the opinions of different researchers. The result is around even likelihood for the validity of some form of DA, and even smaller for "Strong DA", which predicts the end of the world in the coming centuries. We discuss several examples of the validity of the DA in real life as an instrument to test it "experimentally". We also show that DA becomes strongest if it is based on the idea of the "natural reference class" of observers, that is, the observers who know about the DA (i.e. a Self-Referenced DA). Such DA predicts that there is a high probability of a global catastrophe with human extinction in the 21st century, which aligns with many expert opinions of different technological risks.

Highlights:

- There are four main types of DA: future population prediction (Gott's DA), Bayesian update of risks (Carter's DA), the more probable Late Filter (Grace's DA) and the Universal DA.
- Meta-DA treats logical uncertainty about the predictive power of the DA as a probability that DA will work.
- We used a virtual prediction market of scientists to assess the logical uncertainty of the DA, which produced estimation of around 0.4 of its validity.
- The strongest, and thus most important form of DA is the Self-Referenced DA, and for this class "the end" may be as early as middle of the 21th century, though this is not necessarily human extinction.
- Knowledge about the DA could be used to update our global risks prevention strategies by paying more attention to universal risks or by using random prevention strategies.

1 Introduction

The Doomsday Argument (DA) is an attempt to predict the duration of the future existence of human civilization based on our current position in time. The simplest form of DA is based on Copernican mediocrity principle, according to which our position in time is most likely somewhere mid-way between humanity's beginning and end (there are several other types of DA which will be discussed). We approximately know the time of the beginning of civilization and our current position; the DA claims this information may be used to predict the future end of the world.

However, the problem is that such prediction contradicts our expectations about an infinitely long future existence of the human civilization. The predictions could be different (depending on the choice of the referent class and DA subtype), but they typically predict the end soon, from decades to a hundred thousand years from now, while humanity as a space faring civilization could exist for billions of years. So, the problem with DA, which makes it appear paradoxical, is that DA predicts humanity's end sooner than our optimistic predictions about near-infinite human space colonization.

Most scientists feel obliged to disprove the DA; many refutation attempts exist. In this article, we will explore the consequences of the DA being correct.

There are two possible views of the DA: the view of a mathematician, who sees it as one particular solution of an exotic mathematical problem, like the Sleeping Beauty thought experiment (Bostrom 2007), and the view of a physicist or futurologist, who sees the DA as a prediction of the end of the world in the near term. The main difference between these two views is that the physicist is more interested in experimental tests of DA's logic, while the mathematician is more interested in the formal proof. The futurologist is interested in the practical consequences and the ways to use this knowledge to affect the future.

In this article, we will take the point of view of the futurologist, who thinks that there is some chance that the DA is valid, and we will explore the probability of its validity, its consequences, and the ways to counteract and use the DA.

2. Types of DA

2.1. Gott's DA and Laplace rule of succession

The first version of DA which become widely known was described by Gott (Gott III 1993). When he saw the Berlin Wall for first time in the 1960s, he thought that he was probably observing it in a random moment of the wall's existence, and thus something like the Copernican principle of mediocrity was applicable: the wall would continue to exist for approximately the same time it existed before the first observation by Gott.

When the wall fell in 1989, it was around the time of this prediction; this prompted Gott to publish his version of the DA, where he tried to apply the same logic to the existence of the human race. In 1993, he published an article where he stated: "Assuming the Copernican principle that random intelligent observers on earth are not privileged, limits of 0.2 million to 8 million yrs are placed at the 95 percent confidence level on the total longevity of the human species. It is further argued that the odds against our colonizing the Galaxy and surviving to the far future are very long. The argument also explains why intelligent extraterrestrial life has not been detected" (Gott III 1993).

The now moment t_{now} is randomly located between the beginning and the end of the existence of our civilization, which could be expressed, according to Gott, through the random variable *r*:

$$\frac{(t_{now}-t_{begin})}{(t_{end}-t_{begin})}=r\in(0,1) \qquad (1)$$

and as r is random, r belongs to the middle of the interval with the probability *P* for any *P*:

$$0.5(1-P)<r<1-0.5(1-P) \qquad (2)$$

or, by combination with Equation 1, Gott derived Equation 3 for the 95 percent confidence interval:

$$\frac{1}{39}t_{past}<t_{future}<39\,t_{past} \qquad (3)$$

Gott then estimated of the age of the human species at 200 000 years; applying Equation 3, his result was (with a 95 percent confidence level) an expected future end of human civilization between 5100 and 7.8 million years from now.

Gott's DA doesn't rely explicitly on any self-sampling assumptions or changes of the number of observers, as it describes observation of some *external* process at a random moment. This external process may be completely non-sentient, like the sunrise. However, Gott's DA can take observers into account if we ask how observations are distributed during the object's time in existence. For example, if one takes into account that the stream of tourists to the Berlin Wall was constantly growing, then the wall's random observation moment should be shifted toward the end of existence of the wall. The same is true in the case of humanity's existence, as the human population is constantly growing, and observers are distributed unequally.

To account for this bias, it is better to use not the age of civilization, but the *birth rank* of people. In that case, the age of the civilization will be measured in the birth rank of current observers, which is now around 100 billion [ref]. This change, however, shortens the prediction of Gott's DA, as most humans who have ever lived have lived in the past few centuries; thus, it predicts we have only a few centuries more at such population.

Also, not every observation equally counts, as we will discuss later. Gott's argument collapses to the problem of choosing the correct reference class, because if we account for the birth rank, we should say whose rank is actually valid.

Gott's equation for the birth rank of observers is

$$P(N\le Z)=\frac{Z-n}{Z} \qquad (4)$$

where *N* is total number of humans in the world, *Z* is the population number in question, which we may not reach with probability *P*, and *n* is the observer's current birth rank. In particular, with 95 per cent probability, we will never reach 20 times total population:

$$P(N<20n)=\frac{19}{20} \qquad (5)$$

According to Gott's DA logic, since the total human population through all of history to date is estimated to be 100 billion, we will likely not reach 2 trillion.

We could convert this estimation back in time - consider the case that population stabilizes at 10 billion people with a 100-year life expectancy. In that case, a population of 2 trillion people will be reached in just 20 000 years, with 95 per cent probability humanity will go extinct before this moment. Note that this is 400 times sooner prediction than Gott's original prediction of almost

8 million years of human existence. If we assume an even higher future population, including prospective space colonization, we will get even shorter predictions of the duration of humanity.

Interestingly, a similar problem was explored by Laplace at the beginning of the 19th century (Zabell 1989). Laplace asked what the probability is that the sun will rise again, given that it has risen *N* = 6000x365 times before this day, without exception. It is easy to see that the rule of succession is the same rule as Gott's equation if we use it for prediction of when the succession will likely end. (Sandberg wrote about different priors for Sunrise problems.)

Bostrom has pointed out a flaw in Gott's DA: it can be used to predict only events which humanity cannot influence, but this is not true of humanity's "life expectancy" (Bostrom 2013a). Another flaw, according to Bostrom, is that Gott's DA doesn't take into account "the probability of your observation occurring at a time when the phenomenon is taking place may be positively correlated with the duration of the phenomenon". In other words, we are not random observers, but instead we have information about being early in our potential progression as a species.

2.2. Carter's DA: I live in a shorter world

2.2.1. Bayesian update of risk probability based on observer's position in time

Carter invented the anthropic principle in 1973 (Carter 1974). He also came to the idea of DA at the same time but decided not to publish it at that moment, as he thought that presenting the second two weird ideas simultaneously would spoil perception [ref]. His idea of the DA was popularized in Leslie's book "The end of the world. Science and ethics of human extinction" (Leslie 1996).

His idea of the DA is presented in the form of conditional probability: if in the future there is a risk A, how much we should shift our estimation of the risk A because we find ourselves in the time before the event? In other words, if there are two possible futures, let's call them "short world" and "long world", the short one is more probable. For example, the short world is the world in which humanity becomes extinct in the next few centuries, and the long world is the one where it will exist millions of years into the future. The DA favors the idea that humanity is in the short world, as in that case, we will be exactly in the middle of the existence of human civilization. However, if we are in the long world, it will be surprising to find ourselves in so early a moment of its existence.

The simplest version of Carter's equation is just a Bayes theorem, and the full equation for DA can be found in a paper by Bostrom (Bostrom 1997):

$$P(A|E)=\frac{P(E|A)P(A)}{P(E)}=\text{¿}\,\frac{P(A)}{P(A)+\frac{N_1}{N_2}(1-P(A))} \qquad (6)$$

Carter's formula is presented as Bayesian probability update, where known probability A is updated by the fact that we are before the event. Carter argument requires some assumptions, like a deterministic world, as suggested by Leslie, because it has the paradoxical ability to predict the future (Leslie 1996). For example, if an extinction event, like nuclear war, has an *a priori* probability *P*(*A*) = 1 per cent, and there are two possible futures, "Under assumption A has included 50 billion individuals. Under assumption B, humanity has included 5 trillion individuals"[1], when updating, based on the fact that we live so early (*E*) gives *P*(*A*|*E*) = 50.25 per cent, or a 50 times higher probability than the initial 1 per cent.

In some sense, Carter's equation is a binary case of Gott's equation. If we assume that there are only two possible future durations of humanity, say, 200 billion or 200 trillion years, Carter's equation tells us that given our early position, the first outcome is 1000 times more probable than the second, not 0.5 probable, if we assume both outcomes as *a priori* having equal probability. But Carter's equation could also be able to account for the situation when they are not *a priori* equally

[1] see whole calculation in Wikipedia: https://en.wikipedia.org/wiki/Doomsday_argument#cite_note-15

probable, like in the case if we expect just one risk event in future with known *a priori* probability, like a nuclear war.

Carter's equation is based on several assumptions that are not used in Gott's formulation: a) the future human duration of humanity should not be regarded as infinite, or one cannot apply Carter's equation; b) the future of humanity should be deterministically fixed, so we could predict outcomes of future random events, like a nuclear war; c) there are no extraterrestrial intelligences (Bostrom 1997; Ćirković and Milošević-Zdjelar 2003). These three requirements make Carter's equation less convincing. Gott's equation does not have these problems, as it does not make comparison between two outcomes. The deterministic requirement can be avoided by using the DA in a quantum multiverse or in a statistical form, as discussed later.

2.2.2. The Sleeping Beauty problem as a thought experiment illustrating the DA

The mathematical simplification of Carter's model is the so-called "Sleeping Beauty Problem". Sleeping Beauty will be awakening either once (on Monday) or twice (on Monday and Tuesday), with the outcome depending on a fair coin toss. The problem is, if Beauty was awakened on Monday, how should she estimate the probability of being in a short world, in which she awakens only once, or a "long world", in which she will be awakened again.

Here there are two lines of reasoning. One, known as the "halfer" position, states that as the coin is fair, there is a 0.5 probability she is in the longer world, and as in the longer world she is awakened twice, there is only a 0.25 chance that she is in a Monday of the long world. However, there is still a 0.5 chance of being in a Monday of the short world, (i.e. 2 times more likely). So, after awakening on Monday, Sleeping Beauty should expect it to be 2 times more likely that she in the short world and in the long world, which follows the DA: shorter worlds are more probable.

Another line of reasoning is the so-called "thirder" position: in total, there are 3 copies of the Sleeping Beauty (each of which will exist with an equal probability of 0.5, so we may ignore the difference of the "measure of existence" of each), and thus the probability of being one of each of the copies is 1/3. For Monday, the probability is equal for both worlds, so knowing that it is Monday does not update any preexisting probability of being in short world and long world.

The difference between the positions is the way in which the observer should count her own copies. The first is the self-sampling assumption (SSA), which in Bostrom's formulation states "Given the fact that you exist, you should (other things equal) favor hypotheses according to which many observers exist over hypotheses on which few observers exist" (Nick Bostrom and Cirković 2003).

In the case of Sleeping Beauty, she actually exists in only one of the possible worlds, and thus the Beauty divides her "probability mass" between Monday and Tuesday. SSA is known to create a DA.

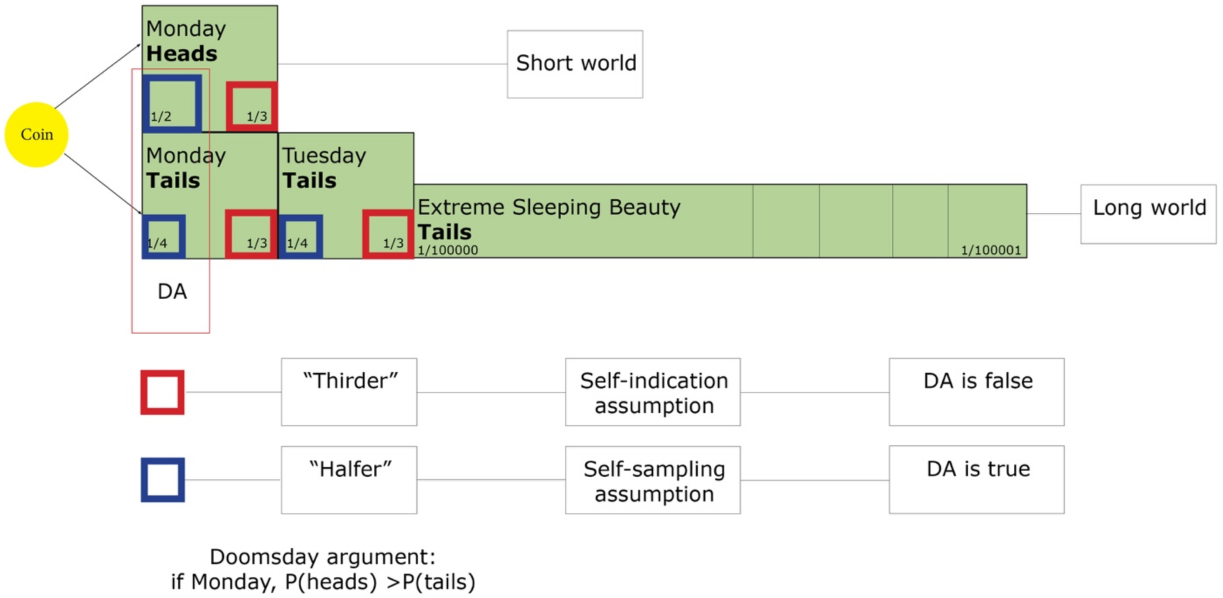
In order to solve the problem, Self-Indication Assumption (SIA) has been described by Bostrom: "All other things equal, an observer should reason as if they are randomly selected from the set of all possible observers" (Bostrom and Cirković 2003). In SIA, the fact that you exist is an argument that you are a member of a larger group, which exactly compensates for the DA's shift in the direction of smaller group; thus, there is no update of the initial probabilities. There is also no reference class problem in SIA, as all possible observers are included—however, there is a question who should be counted as observer, which basically recreates the reference class problem. However, if one assumes a multiverse with infinitely many observers, the SIA loses its predictive power, as all possible observers exist, and the SSA returns. More about this can be found in Appendix 1.

At a deeper level, the question about SSA or SIA is about how the "moment now" appears at any given moment of the historic time. One line of reasoning (similar to SSA) here is that the "moment now" falls randomly in one of several "preexisting slots", which gives us the ability to make some conclusions about the slots' distribution, like the DA. Another line of reasoning is that there is no random tossing of the "now moment" into the time slots, but each time slot already includes all thoughts based only on its local information, and this model is in some sense “updateless” (Yudkowsky 2016). For example, according to updateless logic, any mind who lives in the beginning of 21st century could have thoughts about the DA no matter how long the civilization will exist. This connects DA with the question of the nature of consciousness: if consciousness is a “thing”, like a soul which reincarnates in a body, then it favors SSA. If thoughts are purely mathematical processes, this favors SIA. The unresolved problems of the nature of qualia (Chalmers 1996) cannot help us to decide which theory is true and this question is beyond the scope of this work.

Another important part of the puzzle, which is underlined by the Sleeping Beauty problem, is the nature of “probability”. It can be shown that if we define probability through bets paid for the Beauty by correct guesses, the choice between halfers and thirders’ position depends on if Beauty could accumulate prizes[2]. The “doomsday probability” is also difficult to define and measure, but the necessary level of doomsday prevention efforts may be defined much better than the risk’s probability (Turchin and Denkenberger 2018a).

The Sleeping Beauty problem follows the main pattern of analytic philosophy [ref]: 1. simplify a real-world problem to some thought experiment; 2. Study it extensively; 3. Translate conclusions back to the real-world problem. Steps 1 and 3 may be the source of serious errors. There are many attempts to solve the Sleeping Beauty problem but there is no consensus. There are literally several hundred scientific articles discussing the Sleeping Beauty problem in last 15 years (Google Scholar 2018) that propose many exotic solutions (but only around hundred addressing Doomsday argument explicitly and often without applying to Sleeping Beauty)

Figure 1. The map of Sleeping beauty solutions



2.3. Grace’s DA: Great Filter ahead

As it is discussed above, DA looks like a counterfactual attempt to predict the future of our civilization by somehow learning information about the future. However, we could phrase this differently: given that humanity is a typical civilization (no matter how remote other civilizations may be), what does the DA predict about the fate of the typical civilization?

[2] though many authors claim that betting is the incorrect approach, as we must define “probability” before any decision theory, as discussed on LessWrong https://www.lesswrong.com/posts/aKcy8428zspgSKjYA/sleeping-beauty-resolved-pt-2-identity-and-betting

It has been thought that the SIA solves the DA, but K. Grace created her own DA using SIA by applying the same logic to the space of all possible civilizations (Grace 2010). She asked what is more probable: that the Fermi paradox's Great Filter (GF) is ahead of us (and thus relatively soon, before we become a space-faring civilization), or behind us? She shows that a late GF means a larger number of young civilizations, and as the SIA favors worlds with a larger total number of observers, it also favors a late GF.

In other words, if there are two explanations of the Fermi paradox: 1) where life-supporting planets are very rare, e.g. one in the observable universe, the "Rare Earth" (Ward and Brownlee 2003; Sandberg, Drexler, and Ord 2017); and 2) that there are millions of civilizations in the observable universe but all technological civilizations self-destruct before starting a colonization wave. Suppose, *a priori*, one estimates the probabilities of these two scenarios as equal, when applying the SIA, the probability shifts toward 2, the "GF ahead" hypothesis, with a probability of around 0.999999.

Grace DA assumes that if there is strong past GF, there is weak GF ahead. However, if GF-ahead is a variable, independent of the properties of the universe, like probability of AI goes rogue, in that case Grace DA says less about it. ????

2.4. Universal DA

Grace's DA may be reformulated as an "average age of civilization" argument. We are a typical civilization of average type and we have an average age of all civilizations; in other words, most of the observers in all possible universes exist in human-like young technological civilizations (but not in *Star Wars*-style galactic empires). Does it mean doomsday is inevitable? No; our young age as average may be explained by simulation abundance or future population decline, which will be discussed later.

A similar idea is presented in the article by Knobe, Olum and Vilenkin at al (2006), "Philosophical Implications of Inflationary Cosmology". In contrast to the DA in the spirit of Carter-Leslie, they advance a "Universal DA". Namely, they show that from the fact that we find ourselves in the early stages of humanity, it follows with a high probability, that the set of all people in short-lived civilizations is larger than the set of all people who are in all long-lived civilizations throughout the universe, or, in other words, the number of long-lived civilizations is extremely small.

This again means that the chance for our civilization to become a long-lived one is very small, but it changes the probable cause of human extinction; namely, it will happen not because of some particular reason relevant only to the Earth, but because of some universal cause that could act on all planetary civilizations, even those where the laws of physics are different. They write: "You should not worry especially about the chance that some specific nearby star will become a supernova, but more about the chance that supernovas are more deadly to nearby life than we believe. Many other examples are possible" (Knobe, Olum, and Vilenkin 2006).

Gerig, Olum and Vilenkin (2013) later provided a full mathematical framework which takes into account different prior probability distributions of existential threats in the universal DA. Depending of the type of distribution of short-living and long living civilizations in the universe and the number of threats, their prediction may be either rather mild or strong. They concluded that in the case of uniform prior distribution, "[f]or example, when R = 1 million [ratio of long to short-living civilization members], our civilization's chance of long-term survival is approximately 7%." But in the worst combination of assumptions, their estimation of humanity's chances of survival is vanishingly small, on the order of 10^{-12}.

Note that the same conclusion, that the probability of many short-lived civilizations far exceeds that of many long-lived civilizations, also follows from the discussion about Bostrom's simulation (Bostrom 2003), if simulations are considered short-lived civilizations. The idea that the "GF is ahead" is another representation of the idea that number of short-lived civilizations is overwhelmingly large.

If we compare Grace DA and Vilenkin's universal DA, the main difference is that Grace DA is strongly connected with idea of Fermi paradox and GF, which is known to exist either before us or ahead. This makes Grace DA much stronger, especially given that late GF could happen very soon because of the existential-risks in the 21st century (Bostrom 2002).

2.5. Correlation between different forms of DA

It looks like there are two types of DA – the first takes one's birth rank to calculate doom timing; another ignores individuals, and look only on the properties of the civilization as whole: either its age, as in Universal DA, or its location relative to the GF, as in case of Grace DA.

As we will discuss later, different versions of DA predict different ends of different reference classes, which technically could be explained by different events. For example, the end of human civilization may not correspond to human extinction. Or the end of Homo sapiens may not be the end of humanity if humanity is continued in another form, such as artificial general intelligence (AGI). Or even the extinction of humanity could be reversed later [ref]. Generally, DA variants are better explained by just one event which is the end for all reference classes, that is, by extinction.

Our current opinion is that considering all forms, Grace DA is strongest, as it is based on SIA, which is the only correct form of sampling in the multiverse (see Appendix 2) and also it takes evidence from the Fermi Paradox. Self-referential DA, described in section 5, is also likely true, and while it doesn't predict the global catastrophe per se, combined with Grace DA, it indicates that the catastrophic explanation is most likely.

3. Practical examples of the use of DA logic

As one sees in the section above, there is great uncertainty about the correct type of DA and its validity. In this section, we will check if it is possible to use the DA to get meaningful predictions about already known or verifiable events. This will be an experimental test of DA.

3.1. Measuring the length of the year by the date of birthday

If we apply the mediocrity principle to the date of birth of a person, we could predict that it is most likely an individual's birthday is in the middle of year, not on 1 January or 31 of December, with probability 0.9945. In the author's case, his birthday is the 4th of September, which is close to the middle of the year.

Imagine that we do not know how long the year is, but we could use the date of birth of a random person for this estimation, and the author happens to be chosen as the random person (the reader can use his/her own date of birth). Using Gott's equation, we get a 50 per cent probability that the length of the year is above $2x$, where x is the date of birth in days from the beginning of the year. In my case, my birthday is on the 246th day, and thus the expected length of the year is 492 days, which is an error of 34 per cent compared with the real length of 365. This is rather a good result, if we are interested in the estimation within an order of magnitude.

3.2. Human life expectancy estimation based on my age

We could use the logic above to get a median human life expectancy and estimate my own life expectancy based on this average. This calculation is much more similar to the original DA than the one discussed in the section above, as it is used to predict the moment of death.

I assume that I am a random person and my age is distributed between all ages of all random persons, so there is no causal connection between the moment in my life when I am asking this question and my actual age. This is not exactly true, as it is unlikely that I would ask it in early childhood, say, before age 6, or when I am very old, if I will have dementia, but for the sake of the argument, we will assume a linear distribution of probability of asking this question at any age.

My age now (as of the moment of writing these words) is 44 years old, and using Gott's equation and my age, we calculate that the median human life expectancy is 88 years old with 50 per cent probability. The important point here is that we don't use Gott's equation to directly predict my own age. Instead, we predict the median life expectancy of the members of my reference class. However, my own life expectancy should be assumed to be the same as median life expectancy of the members of my class, as I am a random member of this class. This trick was discussed in the paper by Knobe, Olum and Vikenkin (2006) in which they put forth their universal DA.

3.3. Gott's prediction of Broadway shows' duration

Gott was the first who tried to prove the validity of the DA by using it to predict real world events before they happened. He used it to predict for how long the run of Broadway shows would be based only on the number of previous shows and the assumption that he doesn't know anything about these shows, and measurement of their "age" at a random moment. His predictions of the Broadway shows were consistent with real duration (Gott 1999).

3.4. "End of the galaxy" thought experiment

Earth is located approximately in the middle of the Milky Way Galaxy as measured by its distance from the center of the galaxy. However, we could imagine that in some universe astronomers found that their planet is located around the star most remote from the Galactic center. They could formulate two explanations: (a) there is an unknown physical process their galaxy which prevents life's existence and (b) they are just lucky to be there, as some civilization should be in this location. However, if we take the second type of explanation, we undermine the nature of science which is based on idea of "validity" of experiences, as was mentioned by Bostrom

3.5. DA could be used to predict the future, but uncertainty remains

The suggested above examples show that logic, similar to DA, may be used to predict the future, but this is still may be not enough to validate the use of DA to predict the end of the world. The reasons for it are (Bostrom 2002b):

a) Is our position in the history of the world actually random?
b) Did we correctly account for the change in the number of observers?
c) Have we used some information which was inside the question to reach our answer?
d) What about the difference between the SSA and the SIA - which may be more accurate in cosmological cases versus mundane cases?
e) What if the "end" predicted by DA is not an extinction event?

4. Meta-DA

Many previous attempted solutions to DA look like several pages of Bayesian calculation after which the problem is declared solved, despite obvious false claims or dubious examples provided by the author. An example is Caves' Bayesian rebuttal (Caves 2000), where he suggest the bet "Gott's rule predicts that each dog will survive to twice its present age with probability 1/2. For each of the 6 dogs above 10 years old on the list, I am offering to bet Gott $1,000 US, at odds of 2:1 in his favor, that the dog will not survive to twice its age on 3 December 1999" – but he selects for his bet only dogs who are known to be old, not all random dogs. But DA is a *statistical* argument, which cannot be disproved by a specially constructed unlikely example, in the same way as a speed of a particle in the air cannot replace median temperature, as it is always possible to find a molecule travelling at a speed that would be considered a statistical outlier.

Philpaper.com currently lists 107 articles about DA that suggest many different solutions. There are also many important blogposts about DA on *LessWrong* and *Meteuphoric* blogs, and several books. This makes assessing DA computationally complex for a human, as reading everything written about DA may take months, if not years. The perception of DA may be also biased by the reader’s unwillingness to accept its conclusion or by some semi-undefined preferences, like a final choice between SSA and SIA. All this suggests that before an analysis of DA is finished, one cannot be certain what the conclusion will be, and thus one should be in a state of “logical uncertainty”.

4.1. Logical uncertainty

In recent years, the idea of logical uncertainty has become popular (Garrabrant et al. 2017). An example of this type of uncertainty is about a potentially provable mathematical statement, which, however, is not yet proved or disproved. The work of MIRI has suggested that logical uncertainty should be solved via prediction markets (Garrabrant et al. 2016). In simple cases, logically equal outcomes should be given equal probability, like the 0.1 probability that the n_{th} digit of the irrational number *pi* will be 9.

4.2. Measuring the logical uncertainty of DA

If one looks at the main set of publications about DA, one will see that most scientists are trying to disprove it, and this can be regarded as some form of prediction market, where they put their reputation at stake by taking a stand. Several researchers are explicitly for the DA, including Gott (Gott III 1993), Carter who didn’t publish his findings , but influenced Leslie (Bostrom 2012), Wells (Wells 2009), K.Grace, Vilenkin (Gerig, Olum, and Vilenkin 2013), and Simpson (Simpson 2016), with Bostrom taking an uncertain position. Bostrom has published a full list (Bostrom 2013) but not all of those from Bostrom list have actively published in support of DA; here we include only articles authors, as they are easily accountable is important original contributions. There are more than 10 scientists who have tried to disprove the DA, including Caves (Caves 2000), Olum (Olum 2002), Korb (Korb and Oliver 1998), Sowers (Sowers Jr 2002), Sober (Sober 2003), Dieks (Dieks 1992), Oliver (Korb and Oliver 1998), Monton (Monton 2003), Aranyosi (Aranyosi 2004), Alasdair (Alasdair 2017), and Weintraub (Weintraub 2009).

This kind of informal poll is the first approximation of the prediction market, which was suggested as an instrument for calculating logical uncertainty. Both groups of scientists appear at first glance to be approximately equally large, but the exact number presented here is 8 for DA, and 11 against, which means that there are more scientists trying to disprove DA than to prove it. This implies 8/(8+11) = 0.42 probability of DA being true. Given large uncertainty in adding scientists in the list, this could be approximated as around even credence in DA or against it.

However, there could be a selection bias, as a powerful statement made by prominent scientists may attract less known figures who may try to make their own carrier on disproving the statement, so we should also account for significance of the contributors.

This could be done by weighting using the Hirsh index, I_n, of different scientists. We take median Hirsh indexes of all proponents and divide by the total:

$$L=\frac{\frac{1}{N_d}\sum_{n=1}^{n=N_d} I_n}{\frac{1}{N_{DA}}\sum_{n=1}^{n=N_{DA}} I_n+\frac{1}{N_{nonDA}}\sum_{n=1}^{n=N_{nonDA}} I_n} \quad (7)$$

For the first approximation of the value we will take 20 listed above main scientists (the next step would be to go through all literature on DA in Google Scholar and calculate the value for it). I used Scopus for h-index as it is accessible for all authors, but it typically provides smaller values for h-index than Google Scholar, as it uses different algorithm.

Proponent of DA	**Hirsh index as of 2019, by Scopus**	**Opponent of DA**	**Hirsh index as of 2019**

Leslie	9, link	Caves	50, link
Carter	39, link	Olum	20, link
Gott	38, link	Korb	14, link
Wells	3, link	Sowers	No data: at least 1
Vilenkin	66, link	Sober	28, link
Grace	No Scopus data: at least 3 in GS.	Dieks	16, link
Simpson	15, link	Oliver	10, link
Bostrom	15, link	Monton	9, link
		Aranyosi	2, link
		Alasdair	2, link
		Weintraub	6, link
Total:	188:8=23.5		158:20=7.9

Applying equation (7) to the data from table, we get L = 23.5:31.4= 0.74.

In other words, while there are less articles in support of DA, they are written by more prominent scientists, who has higher cumulative h-index, which result in significant shift in the credence in DA if favor to support it (especially given that most of opponents h-mass comes from Caves, whose interpretation of the Doomsday argument is obviously flawed, as we discussed above.)

In our case, most DA proponents and opponents are prominent scientists, so they have a high Hirsh index, but in other cases, like denial of global climate change, such a metric would provide a significant shift toward more mainstream science. In any case, the metric is not perfect, as it is subject to the Goodhart effect (Manheim and Garrabrant 2018) and other biases. Supporters of DA may also support a different version of DA, a possibility not taken into consideration here.

However, the predictions of the DA themselves are very uncertain, and in most cases, have an uncertainty of at least an order of magnitude in their predictions (most of this uncertainty is based not on the unknown numbers, like my birth rank in humanity, but on question which number should we take and on the probabilistic nature of the DA prediction).

This means that despite all the work done by different scientists, our outside view understanding of the DA’s validity has not changed significantly. And will not change significantly if we publish one more article supporting or questioning DA.

For practical reasons, we may still think that odds in favor of DA or against it is around even.

4.3. Unknowability of DA

In many cases, one cannot know if an extinction catastrophe is happening or not, as at the moment of extinction, there will be no observers. Imagine that at the end of 21st century, a killer plague appears, and it will almost surely wipe out humanity; but a group of scientist survivors will know that the end is near. The question is, could this group use the plague as an evidence that DA was right?

The first scientist might reason: DA predicted the end would be soon, and the end has happened, so this is evidence for DA.

However, the second scientist might say: the plague is just a random event, which was not inevitable. Even if the plague had a 50 per cent probability to happen, another 50 per cent probability was that humanity would survive almost forever, so the plague does not support DA.

The third scientist might say: given how many different dangerous technologies we created, some type of catastrophe was inevitable, and this is an argument for the DA, but not the plague itself as the means of extinction.

4.4. Applying logical uncertainty to the DA's predictions

The application of logical uncertainty to DA seems to be rather straightforward: we should multiply confidence *L* by the DA prediction of the end being soon, *P* (the alternative is that there will be no doom in billions of years from now, 1-P; however, different versions of DA predicts different versions of almost inevitable extinction, form decades to millions of years – and this difference will be for now ignored.)

$$P_{updated} = LP \qquad (8)$$

But it is better to use the following prediction matrix:

- *L*: The end is soon, with probability distribution *P*(*t*)
- 1-*L*: There is no information available about the end (though this fact does not guarantee survival).

Such uncertainty leaves room for hope, but also implies that one should take DA's predictions seriously and invest in the prevention and mitigation of existential risks (Bostrom 2003; Bostrom 2013b; Torres 2016)

In our case, if we take preliminary result from equation (7)

1) DA is false = 0.26.
2) DA is true = 0.74. This could be divided in two groups: (a) weak predictions: original prediction by Gott and rather mild result from Vilenkin's recent article (b) strong prediction, which all versions of Carter's argument as they update small known risks of expected catastrophe, but we independently known that such catastrophes are possible in the 21th century. Thus, 8 main authors divide in two groups of 2 and 6 authors, and if we use it as an evidence for the probability mass which should be assigned to weak and strong predictions, it will be:
   a) mild version of DA is true = 0.74:3=0.246(6); extinction becomes inevitable after thousands or maybe even millions of years from now.
   b) strong version of DA is true = 0.493(3); extinction is inevitable soon, in the next century or a few centuries.

Here we apply logical uncertainty not only to DA itself, but to its two main types, discussed in the next section.

Thus, meta-DA gives us a 0.493≈0.5 probability of human extinction in next few centuries. This estimation is not very surprising given all we know about possible global risks, and approximately the same estimate has been presented by many scientists based on their assessment of global risks without accounting for DA (Rees 2003; Bostrom 2002). (However, many of them they think if we make it through, the number of future human lives is enormous. So even with 1% chance of making it, the expected number of lives way high than DA predicts.)

4.5. Different predictions of different DA versions

Different types of DA and their predictions are presented in Table 1.

Table 1. Different versions of DA and their predictions.

DA type	Reference class	95 per cent timing of "human extinction"
1.1. Gott's DA for humans	*Homo sapiens*, 200 000 years old	4 million years from now
1.2. Gott's DA for civilization	Human civilization, 5000 years old	100 000 years from now
1.3. Gott's DA for human rank	60 billion humans were born (Leslie – according to Oliver	12 000 AD 1.140 trillion humans, or 9120

	and Korb (Korb and Oliver 1998))	years, assuming constant population and life expectancy
1.4.1. Gott's type of DA for those who are able to understand DA, e.g. mathematicians, in years.	Starting from the 18th century, as Bayes and Laplace lived at that time, 200 years ago	6 000 AD 4 000 years from now
1.4.2. Gott's DA for those who able to understand the DA, e.g. mathematicians, in birth rank.	Accounting for growth of the number of mathematicians: probably a million people, mostly in the second half of the 20th century	3 000 AD 20 million more mathematicians, or assuming the same density as the end of 20th century, the number will be reached in 1 000 years
1.5. Gott's DA for those who know the DA, in birth rank and date	First Laplace, then Carter and Gott; now at least 10 000 people know about the DA and think about it. The idea of the DA became widely popular after 2000	In the 24th century, 200 000 more people will think about the DA as an unsolved problem in the next 360 years. (but 50 per cent extinction probability is sooner, in 2036)
2.1. Carter's DA, simple version	If humanity has 1 per cent probability of extinction before 2150 century and 50 billion people before that moment, or 5 trillion people (50 000 years), then applying Bayes' theorem updates this extinction probability to 50 per cent[3]. But it also postulates extinction in 50k years.	21st century? 1 per cent risk of extinction in next 150 years updated to 50 per cent, but extinction within 50 000 years is postulated by the nature of comparative Carter's DA
2.2. Carter's DA, realistic version	10 per cent *a priori* extinction risk, 500 trillion humans after 5 million years	21st century? 10 per cent risk is updated to 99.1 per cent
2.3. Carter's DA, most complicated case	1 per cent risk, 5 billion years from now	21st century? 1 per cent updated to 99.9 per cent
3.1 Grace's DA	If the initial probability of GF ahead is 0.5 (logical uncertainty), and GF power is 1 in 1 000, then Grace DA updates "GF ahead" hypothesis probability to almost 99.9 per cent	21st century?
3.2. Universal DA by Vilenkin	Civilizations, not humans	Very remote, but depends on assumptions of civilization's distribution

[3] this example is in wiki: https://en.wikipedia.org/wiki/Doomsday_argument#The_Doomsday_Argument_as_a_tricky_problem

5. Self-Referenced DA as the strongest form of DA

Following an overview of the main types of DA and creating an estimate of their validity based on the idea of logical uncertainty, here we will try to refine DA, first by strengthening the case for the strong DA over mild DA. Strong DA is a much more serious warning, and in this section, we will show that the DA becomes very strong if we take the correct reference class.

The stronger the DA prediction, the bigger the difference of such predictions from our normal expectations, and more urgent the need for preventative actions, which are not futile, as discussed in section 6. However, the logical uncertainty creates a larger discount for any exact form of DA: ten different forms of DA are presented in Table 1, and *a priori* they all appear equally likely, which means only a 0.05 estimation of validity for any of the concrete forms (multiplied on general estimation of DA validity of 0.5.)

5.1. DA-Doomers as a correct reference class

The DA depends on the reference class of the observers from which one is randomly chosen. If one is randomly chosen from among all animals that existed in the hundreds of millions years before his/her birth, the DA predicts that there will be hundreds of millions of years more before the end of the world (animals), and there is nothing surprising about that. However, it is obvious that I as an observer cannot be chosen randomly from among all animals, as non-human animals are unable to think about DA, and the fact that one is thinking about it strongly constrains the size of the class of observers from among which one is chosen. (We could still use DA on mammalian species id we think of such species as of an external object which is observed at the random moment of its existence, the same way as Gott observed the Berlin wall; however, as hum could cull all mammals during some existential catastrophe, it is not really a random moment.)

It may be suggested that only humans are able to think abstractly, and thus, only humans are in the DA reference class; but obviously, most humans never thought about DA in mathematical terms to the necessary extent or are not interested in the problem (but most humans could understand the logic that if something is lasted a long time, it is likely to last a long time more).

There are two natural solutions to the DA reference class:

1. I am randomly chosen from those who are *able to think about DA* (let’s call this class “mathematicians”).
2. I am randomly chosen from those who *already thinking about DA* (let’s call them DA-Doomers).

The difference between two is rather symbolic, as those who able to think about DA in full extent will start to think about it, as Kant wrote that the thing that is possible in full extent becomes actual (Kant 1781).

The problem is that if I think that I am randomly chosen from all DA-Doomers, we get very strong version of DA, as DA-Doomers appeared only recently and thus the end should be very soon, in just a few decades from now. The first member of DA-Doomers reference class was Carter, in 1973, joined by just a few his friends in the 1980s. (It was rumored that Carter recognized the importance of DA-doomers class and understood that he was first member of it – and thus felt that this “puts” world in danger, as if he s the first in the class, the class is likely to be very short. Anyway, his position was not actually random as he was the first discoverer of the DA).

The real growth of the DA-Doomers started in the 1990s after the idea was published by Gott (Gott III 1993) and Leslie (Leslie 1996) and was widely discussed in the press. The growth continued in the 2000s, because of the appearance of the internet and many publications by Bostrom and others about the DA.

If we assume 1993 as the beginning of a large DA-Doomers reference class, and it is 2018 now (at the moment of writing this text), the age of the DA-Doomers class is 25 years. Then, with 50 per cent probability, the reference class of DA-Doomers will disappear in 2043, according to Gott’s equation! Interestingly, the dates around 2030-2050 appear in many different predictions of the singularity or the end of the world (Korotayev 2018; Turchin and Denkenberger 2018b; Kurzweil 2006).

So, the end of DA-Doomers may be not a global catastrophe but a complete loss of interest in the problem. In next section we will look at the publishing data to learn is the interest to DA is growing or declining.

5.1.2. Could declining interest to DA explain the DA? Experimental data

Hypothesis 1: "I am randomly selected from all people, who know about DA. The number of such people is growing exponentially, form 1980s until now and will continue to grow. Thus, I am currently located only a 1-2 doubling before the end of this class of people, and such "end" could be best explained by a global catastrophe."

To check the hypothesis, I went Google Trends to check the number of times the words "Doomsday argument" is searched. What I found surprised me: the number of searchers is actually declining. The data is noisy, but it looks like the number of searchers declines from the average of 16 a month in 2008 to 7 in 2017.

Wikipedia views data even from 2015 (no early data available) also shows decline around 2 times between 2015 and 2018.

Google scholar analysis is less clear (obviously not exponential growth of the number of articles, but steady growth of mentions which means more scholars know DA):

Google Scholar articles about DA, 1989-1994, 14 articles (hand counted), 30 mentions of DA.

1995-2000: 15 articles, 50 mentions

2001-2006: 24 articles, 100 mentions

2007-2012: 18 articles, 140 mentions

2013-2018: 20 articles, 160 mentions

It shows that the peak of interest to DA by scholars was around 2000, which should not be surprising, as at the time the idea was relatively new. The growth of mentions could be explained by large "historical introductions" in other articles. However, the number of DA-related articles is now growing again.

When I first got the data of declining interest to DA, I suggest that this could explain the DA:

Hypothesis 2: If there will be no more scientists who are interested in DA, the reference class of those who know about DA will end without end of the world.

However, closer examination of the Google Scholar data doesn't support this second hypothesis either: there is a steady influx of new scientists who try to refute or reanalyze the DA. Moreover, the growth of "mentions" shows that the number of scientists who know about DA is growing, but it is growing not exponentially, but more like logarithmically.

Internet access and general growth of population as well as public interest to science could fuel the growth of the number of those who know about DA. On the other hand, lower number of google searchers means that public interest to the topic has declined, may be as there are less mainstream media publications which could fuel such interest or doomsday media paranoia, like in 2012, which could easily be observed as a spike of searchers around 2012.

The data could be explained if we suggest that less members of public but more scientists now know about DA – and the question is interesting not from sociological perspective, but in order to understand how the reference class of DA-aware observers is changing.

It seems that the correct reference class will be the scientists, not public, as the fact that I am writing this post (and had long detailed interest to DA before) makes me closer to the scientists' reference class.

For scientists, we have two sub-classes: those who know about DA, and those who try to make new contributions by writing articles. The difference is that one is growing and the other is not.

Both hypotheses are false: the hypothesis that the interest to DA is exponentially growing, and the one that the number of those who understand DA is exponentially declining: so, there is no end-very-soon, nor DA's easy refutation.

However, using the reference class of those who know about DA still imply that the end is likely in 21 century. There are currently around 100 scientific articles about DA, and with 50 per cent probability (according to the Gott's version of DA) there will be no more than total of 200 (which at current speed will happen in around 30 years, or 2049) and with 95 per cent – no more than 1000 articles (which will happen at current speed of publishing at 270 years). But such end could mean not a global catastrophe but complete loss of interest in DA.

5.2. Meta-DA-Doomers

Carter, when he discovered the DA in 1973, was worried that he was the first and the only one who knew about the DA, and thus, his reference class was very small, as it consisted of just one person for a few days, and this could mean the end was very soon. One may think that his thoughts were unjustified at the moment, but we now have additional knowledge: that the catastrophe didn't happen.

However, if one identifies the class of those who think about the DA as the correct reference class, we immediately create a new reference class: those who know about the correct DA class—a class from which I am randomly chosen. This class is even smaller, and probably includes only a few people, including me (and perhaps several other people, maybe, including Carter).

We will call the class of observers, who a) knows about the DA; b) think that the correct reference class = "those who know about the DA", – as a class of "Meta-DA-Doomers". It may seem that there could appear an infinite regression of meta-meta levels, but if anyone jumps to the meta level s/he also understands the possibility of additional "meta jumps", and thus there are no special meta-meta level classes of observers.

There should be noted that I discovered the meta-DA idea as early as 2007—and the world has not yet ended, between that moment and the writing of this paper in 2018. For the DA-Doomers, only the moment of discovery of the idea should be counted as random, not the current moment, which is always later.

5.3. DA-Doomers self-refutation

However, this DA-Doomers setup is self-rebutting, as DA-doomer reference class includes only those who are *surprised* by their earlier position, and those who will live later will not be surprised (but may know about DA as a historical fact), so they will have a different cognitive process about the DA and will be not members of our reference class.

But how long one should be surprised by his early position depends of the type of DA and expectations about humanity future. For example, if civilization is expected to exist for billions of years, the ones who lived in first millions of its existence may be still surprised.

5.4. DA predicts the end of the reference class, not the end of the world

DA doesn't say anything specifically about *how* the reference class will end, it just says that it *will* end. We assume that if the class of humans ends it will mean a global catastrophe. But the class of humans may end in different ways.

Maybe humans will be replaced by other sentient beings, like cyborgs or our biological descendants. Perhaps they all will merge into one superintelligence mind.

We could formulate the following principle: *For each reference class, DA predicts its own end*. For example, for "knowledge about the DA", the DA predicts the end of the knowledge, and for human biological beings' births, the DA predicts the end of such births. The interpretation of such an event as a global catastrophe is only an interpretation, which only in some cases seems probable.

For example, if we regard the class of DA-Doomers, the end of this class may mean not the extinction of humanity, but just the fact that a well-known and obvious refutation of the DA is included in its Wiki article, so anyone who becomes interested in the DA will immediately learn that it is false. Or after some point in time, observers will just stop being surprised that they are early in human history.

6. Cheating the DA and using the DA

6.1. Bostrom's UN++ for escaping smaller catastrophes

The DA may be regarded as something like an ominous witches' spell, which one cannot do anything with. But humans are able to find practical uses for many potentially dangerous things, like nuclear energy. Bostrom has suggested several thought experiments, demonstrating how the DA could be turned into something that looks like magic. One example is the "Adam and Eve experiment", in which Adam uses DA-logic as a proof that contraception is not needed as large population is unlikely (Nick Bostrom 2001).

Another idea is his "UN++ thought experiment" (Bostrom 2001). In the future, a powerful UN appears. It learns that some bad thing (but not an extinction event) could happen soon. Its members decide that if such a thing happens, they will increase the global population 10 times—since such a population increase will make the earlier timing of such an event less likely, and according to DA-logic, this translates to lowering the probability of the bad event by 10 times.

In the setup of this experiment the commitment to increase the future population makes the current earlier position less likely, and thus makes precondition of the commitment also less likely – which counterfactually could be used to manipulate probability of this condition – or to disprove DA as absurd. From causal decision theory (CDT) perspective, it is absurd, but the same CDT recommend two boxing in the Newcomb problem, which is a losing strategy. In some sense, DA is Omega from the Newcomb problem, which is able predict my future choices.

If we use Updateless decision theory, we should make choices which makes most agents of our type win. This imply one-boxing in Newcomb problem. But such winning strategy can't be proved causally (by opening the second box). The same way it is impossible to find any causal mechanism which will prove that UN++ experiment (or other DA-relate manipulations) will work.

Another type of manipulation of probabilities by changing the observer numbers is described in Yudkowsky's "Anthropic trilemma" (Yudkowsky 2009). Given the improbable setup and uncertainty that the DA will work at all, attempts to manipulate probabilities, such as the UN++ suggestion, do not appear to have practical applications.

6.2. Escaping the DA: The timer is reset in simulations

If some civilization takes the DA seriously, it may try to escape it. One way to escape the DA is to "forget" the actual time position of the observer. One way to forget is to create many simulations of the past, in which the agent does not know her/his actual birth order as she-he doesn't know the number of simulations. For example, if one watches a movie about life in Ancient Rome, one may temporarily forget one's actual position in time in the 21st century.

While it is unlikely that creating many past simulations could be motivated just by a desire to cheat the DA, as it assumes too-high credence in the DA, it could be an additional bonus for a civilization which has decided to create past simulations for other reasons. Again, it works only if we one-box in Newcomb-like problems.

6.3. Smaller population

DA predicts the end of the world is soon based on the assumption that the human population will continue to be large. If the population becomes small again, on the order of a few million people, the probability is that the predicted doomsday will be postponed, and will happen not in next thousand years, but in the next million years, which is nothing surprising for an ordinary species, but surprising in light of the prediction that humanity will be a technological, spacefaring civilization.

One way of lowering population is coalescence of consciousnesses, such as if consciousnesses are merged into one superintelligent AI, maybe via some form of Neuralink [ref] or neuroweb [ref]. Or if humans are uploaded, they could live in simulations with different clock stamps ("birth orders"), as described in section 6.2.

6.4. The "Cheating Death in Damascus" solution to the DA

Turchin has already discussed this idea as an instrument for escaping the GF of the Fermi paradox (Turchin 2018a; Soares and Levinstein 2017). If there is something that is killing everybody, it seems rational to try random strategies, as all rational ways to escape will have killed previous contenders. But the idea of attempting random strategies may be a killer in itself, so at first, one should toss a coin and decide whether to try a random strategy or the purely rational strategy with the highest expected payoff.

6.5. Strategy for global risk prevention in light of DA

In one of the most elegant forms of DA, "universal DA" by Vilenkin et al. (Knobe, Olum, and Vilenkin 2006), it is underlined that if DA is true, this fact has important practical consequences. That is, that we should search for universal ways how civilizations become extinct, not anything specific to Earth's history. Such possible universal mechanisms include:

1) Artificial intelligence (Bostrom 2014).

2) Accelerated growth of technology and availability of dangerous technologies to smaller and smaller groups (Turchin, Green, and Denkenberger 2017).

3) Population increase and resource depletion (Meadows, Randers, and Meadows 2004).

4) Increase of chaos in more complex systems. Complex systems could become chaotic, that is, it is impossible to predict their future, and thus they could have sudden collapses. (Куркина 2013; Tainter 1990).

6.6. Choosing the best strategy to escape DA's power

Now we need to assess the best way to evade doomsday from those listed above, which include

- the prevention of the universal global risks for all civilizations,
- using random strategies (which is the opposite strategy to preventing universal risks)
- changes in way we calculate our rank, including simulations or smaller population (which is somewhat like smoking lesion decision theory problem (Egan 2007)).
- more research in DA (which may be not useful as it would only increase the number of DA-doomers, and thus cancels explanation of DA through the loss of interest to it).

From all these, prevention of universal sources of risks appears to be the most reasonable approach, as it causally increases survivability even if DA is false. Updateless decision theory would recommend us to do the same.

7. Discussion

My current opinion is that from all forms, Grace DA is strongest, as it is based on SIA, which is the only correct form of sampling in the multiverse (see Appendix 2) and also it takes evidence from the Fermi Paradox. Self-referential DA, described in section 5, is also likely to be true, and while it does not predict the global catastrophe per se, combined with Grace DA, it indicates that the catastrophic explanation is most likely.

8. Conclusion. Living in the middle of the world

In this article, we explored the controversial probabilistic argument called "The Doomsday Argument". Based on the theoretical arguments and evidence from similar situations discussed in the article, some form of DA is likely to be true (most likely, in the form of universal DA of Vilenkin et al or Grace DA).

However, some uncertainty remains. To account for this uncertainty, we created the meta-DA, that is, the probability estimation that some form of DA may be true. Meta-DA predictions do not contradict what we already know about the future: that there are high risks of human extinction but also there is a chance for humanity to survive.

Meta-DA prevents DA from being universal unescapable killer: if there is a probability that DA is false, then attempts to survive it are not futile.

These chances of survival could be increased if we use one or more of several instruments: prevention of universal global risks for all civilizations, using random strategies (which is the opposite strategy to preventing universal risks of highest expected value), or changes in way we calculate our rank. From all these, prevention of universal sources of risks is likely the most promising approach, as it causally increases survivability even if DA is false.

Appendix 1. Other interesting forms of DA

1. Reverse DA: there will be stability in the short-term

Imagine, that you are waiting for a bus, and you know that the last one was 40 minutes ago. What is the probability that it will appear in the next 1 second? This problem is similar to the Laplace Sun Rise problem, and using his equation we could get that chances are below 1:2400. In general, this means, that if we observe some process in a random moment, the chances that it will continue for some time, smaller than the age of the process, are high. The longer is the process, the more probable is that it is stable and will continue even longer.

This could be used as a general counterargument about a prediction that something will end or happen soon. For example, Moore's law is more than 50 years old, and thus it is unlikely that it will end in the next year, and there is only 10 per cent chance that it will end in the next 5 years from the outside view.

We could call this "reverse DA": *it is very unlikely that the end is very soon*. It is in fact was described in the first Gott's article as 1/39t after now when the catastrophe is unlikely (Gott III 1993).

2. DA and Simulation Argument: you are in a simulation and it will be turned off soon

The simulation argument (SA) claims that at least one of three alternatives is true: A) we live in the simulation B) future AI will not be interested in the creation of simulations C) human extinction will happen before superintelligent AI creation (N. Bostrom 2003). It was suggested that DA and SA cancels each other (Aranyosi 2004), or that SA is right but DA is wrong (Lewis 2013).

Firstly, we have to patch SA. SA should not be used to predict the future of only our civilization, as there is some form of circularity. It should be correctly applied to all possible civilizations, because even non-human alien civilizations could model human civilization just as an experiment, the same way as universal DA by Vilenkin (Knobe, Olum, and Vilenkin 2006).

So, SA must be much stronger: At least one of three alternatives is true A) we live in the simulation B) *no alien superintelligence exist in multiverse which is interesting in creation of simulation of other planets*. C) All civilization goes extinct on the early stage. (B) seems *a priori* very unlikely as AIs in the universe will have convergent goal to create simulations of other planets as this will help the AI to numerically solve Fermi Paradox. It is unlikely that all of alien AIs will be so ethical that they decide not to model aliens to minimize suffering (and humans are aliens for them). This also makes variant C much stronger – "all possible civilizations go extinct before they reach ability to create simulations of other civilizations". Both patched B and C are a priori much more unlikely than original claim, which makes variant A – we are in the simulation – even stronger; or we live in very pessimistic universe (and there are no other universes) where current human civilization is maximum possible level of the technological development.

DA could be applied inside a simulation only if we could measure some linear parameter of the duration of our existence. In the general case, people in the simulation do not know how long their simulation has existed: maybe it was created just yesterday with clocks saying that it is existed for 14 billion years. The main point of simulation is that does not simulate all the past, but only a part of time. This means that the clocks in simulations are generally shifted to smaller past running times, and any DA prediction based on the clock reading also should be shortened (never extended).

Thus, being in simulation makes DA stronger, as simulations run shorter time. For example, if one opens a door into a cinema at a random moment, one will find his/herself in a random moment of a ~2 hours movie, and may conclude that it will end in 1 hour, despite the internal clock of the movie showing years of a character's childhood.

If one has access to the correct time of how long the simulation has been running, one will be able to apply DA to learn how long it will most likely run until its turnoff.

SA could be tested the same way as we tested DA above in the toy real life examples. For example, most time we see an expensive object, it is not the real object, but either photo, video, movie, night dream, day dream, computer game, etc. Many people already spent much time in some form of simulations, and their quality and relative duration is growing.

DA and SA combined predicts that most human observer-moments are in simulations, which are just short running simulations of some civilization past. This will be exactly the case for alien AI simulating millions of possible civilizations in order to solve the Fermi paradox, as the AI will simulate only a short historic period, probably corresponding to our 20th and 21st centuries, when most decisions about global risks prevention will be made (They would also be simulating abiogenesis and multicellularity as important events, but there are no observers there, so it will not affect our estimations). If most simulations are just computer games for some advanced beings, they could be also rather short as the game will be concentrated only on most interesting historical moments.

3. DA implication for SETI: human civilization is typical

From the Copernican mediocrity principle, it also follows that human civilization is typical, that is, it is the member of the one of the biggest classes of such civilizations. Surely, if some civilizations have very few specimens (like thinking ocean Solaris (Lem 1970)), they still could dominate by number of civilizations, but not by the number of specimen.

This means that, most likely, other civilizations with which we could contact, are either the same type as ours, or consist of much smaller number of independent agents (which could be one superintelligent AI) – but civilizations consisting of trillions sentient beings are untypical and unlikely to be found.

Typicality of the human civilization increases the chances of the mutual understanding with aliens, but also increases chances that humanity falls victim of the SETI-attack (Turchin 2018b), a malicious SETI message aimed at self-replication which includes a description of a dangerous AI program.

4. DA as argument against superintelligence: human mind is average

In the article (Pereira 2017) suggested the Super Strong Self Sampling Assumption that one should find oneself not only randomly taken from all observer moments, but weighted according the “size” of consciousness. Pereira conclude that this explains why a person having these thoughts is not an animal, despite animals’ their numbers being overwhelmingly larger than human observers. He then concludes that “superconscious” AIs are very unlikely, so human consciousness size is typical. However, using DA-Doomers reference class gives a different explanation to the question why one thinking about this is not an animal, as one should count only from all beings who are capable to think about anthropics and DA. But an interesting point in the article is that the smallest Boltzmann Brains (randomly occurring agglomerations of particles (Carroll 2017)) are much more probable than more complex brains of this type.

This mediocrity logic may be applied not only to the “size of consciousness”, but to the mind size. I am not surprised that my IQ is somewhere between lowest and highest of all humans. But if we extrapolate this logic, we would find that superintelligent minds are not dominating minds by the number of observer-moments in the multiverse. It basically means that superintelligent minds of human architecture are impossible, or at least very rare: there could be other superintelligent minds, which doesn’t have such internal structure as “observer-moments”, as they could use completely different optimization process as the main algorithm.

Some forms of superintelligent minds could be possible if they don’t have observer-moments, for example, the process of natural evolution could be described as superintelligent optimization process, but surely it doesn’t have complex observer-moments or any other agential properties.

But if each superintelligent AI creates numerous past simulations with trillions of human-like observers, then low-intelligence observers would dominate.

5. Regression to the mean and the agents of entropy

Strugatsky suggested in their novel “Billion years before the end of the world” (Strugatsky and Boris 2014) another form of DA: that as our low entropy region of the universe is a fluctuation, it is unlikely that such a low entropy region will continue to grow, and it has to return to the chaotic state. But, according to the plot, human science increases the neg-entropy, and thus there should be tendency in the world to counteract such growth of order. In the plot, agents of chaos try to stop the work of several scientists.

The story should be viewed as example of “DA magic” similar to the “Adam and Eve” thought experiment by Bostrom (Bostrom 2001), where Eve cannot conceive a child as this will increase the future Earth population and will make Adam unlikely to be the first person. The story is designed to look absurd, but absurdity is already in its premise: a mythological story and rational knowledge are mixed in the way where they never can mix in reality. For example, if Adam were to know about Fermi paradox’s GF, he may think that Eve’s inability to conceive is the GF, and it could be true for many Adam’s in different worlds. We *a posteriori* know that Adam would not have been right in our world, but we do not know in how many worlds he would have been right.

If we remove the entertainment narrative from the Strugatsky novel plot, the main idea is akin to the one that was described in the previous section of the article. This idea is that very complex structures are unlikely in the Universe, and that is why we could observe forces which prevent such systems from appearing. For example, if I try to become richest person in the world, there will be many reasons which will prevent this from happening (scammers, law enforcement officers, etc), because to be the richest is a very unlikely event.

In other words, the unexpected difficulty of a task may be perceived as a force preventing it to be solved.

6. Our place in the universe and the Fermi paradox solution

There were attempts to use our position in time in the history of the universe to get something meaningful about the Fermi Paradox. It was found that there is nothing surprising in our position, if it is counted by the number of stars which have been and will be ever born: we are in the middle [ref].

This may be a counterargument to just one particular scenario: that many civilizations appear and quickly burn all available resources (including habitable planets) (Hanson 1998). In that case, we should find ourselves unexpectedly early, as resources needed for our appearance was not use by any other civilization. If we are the only one civilization in the observable universe, as it is assumed by the Rare Earth theory (Ward and Brownlee 2003), this is completely in line with the observation (no ETI), and also if all civilizations kill themselves at technological stage.

7. DA as an argument against immortality

It was suggested that immortality is impossible, or I should find myself already having infinitely high age (Leslie 2008). But DA does not work this way. The moment when I observed something the *first time* should be random.

However, the same logic is applicable to calculating median human age: given my current age of 45 I can predict based on Gott’s equation, that median human life expectancy is like 90, which is very close to actual result – and contradicts expectations about immortality, so it is the same DA, but on personal level.

The main difference form DA is that in original DA the moment of birth is counted, which is fixed and completely independent of my mind process, and in here we count the moment of “now” distribution along the line of life. But not any “now” moment should be counted, but only those there I surprised about my early position, and thus the setup is very sensitive to my thought process, the same way as in described above meta-DA. This makes the argument weaker: maybe when I will be immortal, I will be less surprised about my position, so it will be just different group of observer-moments from which I am “randomly” chosen.

8. DA as instrument to predict AI timing

One of the most important problems of future prediction is prediction of “AI timing”, that is the time of creating AGI. Humanity lives in a period after such research has started but before it comes to fruition. One could say that AI research started in 1956 at Dartmouth workshop [ref], and I wrote first time about using random observation data of AI at 2016, 60 years later – I use my own data as it based on the logic of DA: we use our position as just one sample and assume that it is random. This predicts that AI will appear (or research end for other reasons) with 50 per cent probability in the next 60 years (2076). This prediction surprisingly not differs much from expert poll by Grace (Grace 2017), which gives 2062 median AI timing.

The same logic could be used for predicting nuclear war (assuming that the moment of first writing of the text in 2018 – but not editing – is random moment relative to the observed process).

But the difference is that nukes were actually used in combat in 1945. This means that the next military use would be before 2091 with 50 per cent probability, or around 0.68 percent a year, which is close to estimations by Baum (Barrett, Baum, and Hostetler 2013).

9. DA in multiverse; DA and quantum immortality

In the Everett interpretation of the quantum mechanics, the world is constantly branching, and this could be interpreted as the growth of the number of the observers, which would grow every second many orders of magnitude with each molecular collision (Wallace 2012). If we think that each observer-moment is randomly selected from the all observer-moments (as was suggested by Bostrom's principle of SSSA (Bostrom 2013)), then the biggest number of observer-moments will appear just before the end of world, which could be something like false vacuum decay which immediately kills all observers on Earth (Wilson 2015). If this is true, I should find myself in the last second of my life. This may be emotionally disturbing but does not have any observable consequences and is completely compensated by "quantum immortality": the fact that because of quantum branching there will be always a timeline where the observer will survive (Turchin 2018c).

To invalidate such a prediction, it was suggested to account for the "weight" of each observer, which also called "measure" (and from the mathematics point of view, quantum mechanics is the probability density of each branch). As the world is splitting, the measure is also splitting between new worlds, so the total measure of all branches is not changing, but the measure of each new branch is diminishing. In this case, DA will not move the observers closer to the end of the world in the quantum world.

The splitting world of quantum multiverse also creates the effect known as "quantum immortality", where for any observer exists one timeline where s/he never dies. This could also compensate observable consequence of "next second false vacuum decay", as some tiny fraction of observers will exist in the universe where the false vacuum decay has not happened.

However, the same logic is applicable even to the classical universe. Imagine that the number of observers in the universe is growing, as it is becoming more hospitable to the appearance of new civilizations or the existing civilizations proliferate through the universe. In that case, any observer is more likely to find her/himself closer to the end of universe, as it will be the period of maximum observers' densities. But this conclusion depends on the actual distribution of minds in the universe, which may slowly decay in the heat death scenario.

Based on the material above, we could distinguish weak DA: maybe we live near the end of the world – and Strong DA: intelligence appears only in the world which is near a final catastrophe.

Analogous to the anthropic principle, we could suggest weak and strong DA:

1. Weak DA: Humanity is likely to not exist as long as it may hope to.
2. Strong DA. Intelligence appears only in the world at the verge of a global catastrophe. Strong DA may be the case in the world with powerful anthropic shadow, where the next catastrophe has started to happen, and its harbingers create environmental instability which require capability to quickly adapt the changing environment. For example, periodic ice ages created a constantly changing environment where universal intelligence was a better adaptation than any specialized drives or "feathers". However, such ice ages themselves imply future climatic catastrophe (methane clathrate gun hypothesis [ref]).

10. Anthropic shadow and fragility of our environment

Bostrom et al suggested the idea of *anthropic shadow* (Ćirković, Sandberg, and Bostrom 2010), that is something like survivorship bias in the estimation of frequency of natural catastrophes in the world. Metaphorically, we could say that anthropic principle defended us against any past catastrophes, like extremely large asteroid impacts and false vacuum decay. Bostrom showed that universe scale catastrophes, such as false vacuum decay, are no more often than once in 1 billion years even if we account for possibility of anthropic shadow.

But in the case of smaller local catastrophes like asteroids, supervolcanos and superflares, the same limitations may be much shorter. Moreover, some of such catastrophes may be long overdue. There are two types of natural catastrophe: truly random (asteroids) and cyclic (comets and may be some supervolcanos). For cyclic catastrophes, their energy is accumulating before each of cataclysmic events.

If we apply Bostrom's estimation of survivorship bias to earthly natural catastrophes (Tegmark and Bostrom 2005), they would increase their probability no more than around 10 times, which still seems relatively safe, as the largest global catastrophic volcanic eruption happened hundreds millions years ago during the PT boundary.

However, if we live in the period of unusually long pause between a cyclic natural catastrophe, this means increased fragility of our environment to small impacts, and "environment" here means all needed conditions for our existence, starting form stability of vacuum and Sun, and up to atmosphere composition. While actual volcanology is more complex, the simplified example of logic above would suggest that, if the magmatic chamber of a supervolcano is already filled and under pressure, a small impact like geothermal drilling could provoke the eruption earlier, and as a result large volcanic winter will end our civilization.

Also, the environment could become less stable before the catastrophe. This is especially true for climate, which could be long overdue to transition to another stable state (either snowball earth or runaway global warming). Instability of climate is already observed in the form of ice ages, which may be surprisingly helpful for human general intelligence evolution [ref]. Most animals live in the stable environments using hardwired behavior patterns. But humans have to constantly adapt to different environments and the ways of feeding (scavenging, hunting, agriculture), which required high ability to learn and helped to evolve "universal learner".

In other words, general intelligence is the better adaptation during periods of changing environments, which are more probable closer to the global catastrophe. Thus, it is not surprising that we find ourselves closer to the end of the world.

In the case of Earth, the biggest fragility is anthropogenic global warming, which potential we could underestimate as it is long overdue because of accumulation of methane hydrates in Arctic. These hydrates could create a rapid global warming event through positive feedback loop, as methane is a very powerful greenhouse gas and warming is occurring faster in the Arctic.

11. DA in cosmology

One could assume that we live in a perfectly tuned universe optimized for maximum number of observers. However, most observers may live in not so perfectly fine-tuned universes, if the tuning is happening in multidimensional parametric space, because the volume of such space is growing as power of *n*, where *n* is number of the parameters. This is analogous to the fact that the majority of the Sun's mass is not in its core, despite the fact that the core has highest density.

This might mean that our universe is less fine-tuned than it could be, and this has two consequences:

1) Civilizations are much rarer in space than they could be, so we are more likely to be alone in the observable universe.
2) The universe is less stable and life friendly than it could be, and large catastrophic life-sterilization events, like gamma-ray bursts are more likely.

12. DA and Fermi paradox

Aside from the conclusion that the GF is ahead, there are several other Fermi related considerations.

A recent article suggested that the natural solution to the Fermi Paradox is that the first civilization is capable to prevent existence of any other new civilization, the same way as first life in Earth made impossible appearing of other types of life from primordial soup (Berezin 2018). There are different ways how such "prevention" may happen, including the start false vacuum decay, wave of space colonizing or Berserker probes (those they destroy other civilizations at a certain level of sophistication [ref]), or launching effective SETI-attack.

The reasoning is similar to Bostrom's thought experiment about Adam and Eve [ref], as being the first in any list implies that it is unlikely that there are many members in the list.

In other words, if we are a typical civilization, it is more typical for any civilization to find itself alone (at an early stage of development).

13. Catastrophe types predicted by DA

Universal DA predicts that not only humanity will become extinct soon, but most other civilizations will be short lived, maybe even in universes that have different physical laws – because as we are typical, we live in the typical universe by the number of observers. (But if there is another type of civilizations with only a few observers in any moment of their existence, they are not covered by this universal DA logic. However, in the sense of DA-Doomers referential class, our civilization is very small: only a few thousand people at every moment have understanding of DA. Completely non-human form of intelligence, like – fictional example – "thinking ocean" Solaris – are not included in this logic and could be abandoned in the universe: for example, biological evolution is powerful optimization process, but doesn't have "observers".)

If all civilizations are going extinct, this means that there should be some universal cause of civilization's extinction, which cause civilizations to go extinct shortly after they discover DA, but it could not be anything like nuclear war, as this may depend on local availability of nuclear materials. The universal cause is unlikely to be AI, as at least some civilizations may be able to successfully ban its creation. One possible universal way of extinction is a complexity crisis, that is, the growth of complexity exponentially increases the number of possible global risks. Another possibility is the increase of the number of possible bad agents and simultaneously limits ability to predict and manage the future. Vilenkin suggest that his universal DA means that we should pay more attention on the universal types of catastrophes, not Earth-specific, as we could underestimate them [ref].

14. Different forms of DA may be used to predict different forms of the end for different reference classes

A difficult question is: could different forms of DA be true simultaneously? It seems that one cannot use Gott's DA result as prior probability in Carter's DA, as it would mean double use of the same information.

However, as one is the member of the several reference classes, one could estimate the end of each of them based on my position. For example, animals existed for around 400 millions years, and based on Gott's DA, applied to the class of animals, they could exist hundreds millions years more, which completely unsurprising based on the estimation of the future habitability of Earth which is estimated in 100-1000 millions years from now. Some members of the genus *Homo* may exist for several million years based on the same logic.

If one accepts such a form of multilevel prediction, then the DA predicts that future catastrophe will decimate a) large human population b) most bright minds capable to think about DA, but some humans and many animals may continue to exist. So, it is more like civilizational collapse than a catastrophe with a black hole eating the whole Earth.

15. The number of the past civilizations on Earth

Based on known archaeological data, we are the first technological and symbol-using civilization on Earth (but not the first tool-using species). This leads to an analogy that fits the Fermi Paradox: Why are we the first civilisation on Earth? For example, flight was invented by evolution independently several times. We could imagine that on our planet, many civilisations appeared and also became extinct, and based on mediocre principles, we should be somewhere in the middle. For example, if 10 civilisations appeared, we have only a 10 per cent chance of being the first one.

The fact that we are the first such civilisation has strong predictive power about our expected future: it lowers the probability that there will be any other civilisations on Earth, including non-humans or even a restarting of human civilisation from scratch. It is because, if there will be many civilizations, we should not find ourselves to be the first one (It is some form of DA, the same logic is used in Bostrom's article “Adam and Eve” (Bostrom 2001)).

If humanity is the only civilisation to exist in the history of the Earth, then it will probably become extinct (if it goes extinct at all) not in a mild way, but rather in a way which will prevent any other civilisations from appearing. This means higher probability of future (man-made) catastrophes which will not only end the human civilization, but also prevent any existence of any other civilisations on Earth.

Such catastrophes would kill most multicellular life. Nuclear war or pandemic is not that type of a catastrophe. The catastrophe must be really huge: such as irreversible global warming [ref], grey goo [ref] or black hole in a collider [ref].

Appendix 2: Solving SIA and SSA problem in the universe where all possible observers exist

Many thought experiments with observers are designed in the way as if there are not any other observers in the universe. For example, in the Presumptuous Philosopher thought experiment, it is assumed that there are either a trillion, or trillion-trillion observers in the universe, and no more. However, the contradicts the popular idea of the *multiverse* which implies existence of infinite number of observers, and in which all possible observers actually exist.

In this multiverse situation, an observer cannot use the fact that s/he exists as an argument for anything, but this does not prevent him/her from using some form of DA. Because any random variable s/he observes is still random (like day of his/her birthday), and thus s/he most likely observes it in the middle of the interval. Bostrom wrote about the implications of existence of all possible observers to the possibility of science: some observers should be more probable than other (may be by having higher measure), or our observation will be as random as of Boltzmann brains (Bostrom 2002b).

Self-indication assumption (SSI) could be best explained as that I am randomly chosen from all instances which could create my current experience.

Self-sampling assumptions (SSA) is that I am randomly taken from the members of my reference class.

SSI is equal to the *ultimate reference class* by Almond.

In SIA, the real reference class is "the class of observers who is subjectively indistinguishable from me" – and that is why SIA doesn't depend on any other reference classes which I could be a member. However, it doesn't exclude the use of SSA logic for SSA-related conclusions.

An example of SSA logic: I am a member of a class of people who was born between equator and a pole of Earth, and by the fact of my birth I was randomly selected from this class. Thus, the place of my birth should be rather randomly (but accounting for different population densities) selected between equator and pole, and unlikely to be exactly on the equator or on the pole. I was born at 55 latitude, so SSA logic work in predicting my latitude of birth.

I could be a member of many different SSA-classes and for each of them make independent predictions about my position in them.

In SIA the class of "subjectively indistinguishable" my copies could be also not very exact. Different interpretation of such class is:

1) Everybody is me who have the same thought process as me now. There could be a lot of them, even on Earth. (And such class is used in timeless decision theory, there a thought process is counted independently of any other aspects of personality if such aspects do not affect the thought process). For example, in DA-doomers class is such class.

2) Everybody, who has the same total sum of all visual (and other) experiences as me, even despite the fact that I will not be able to account for all differences as they are too small to account.

2a) Everybody whose observations are indistinguishable from mine from inside, in other words, some pixels could be different but I am not mentioning the difference.

3) Everybody who has exactly the same brain as me. This class many orders of magnitude more rare than (2), as the same experience could be generated by different brains.

I think that "true" SIA class is somewhere between (1) and (2) – or more likely, there is no "true SIA class", the same way as there is no true SSA-class, and different types of SIA could be used to answer different questions.

SSA also is defined as choosing from *actually existing* observers, which become controversial, if we speak about future events – that is why Leslie in his book (Leslie 1996) had to spend a lot to prove that future events actually exist or at least strongly causally defined by current events, which seems in contradiction with quantum mechanics.

SIA is applied to all possible observers, and uses their probabilities as weights.

In the Everetian setup where all possible observers do actually exist, this difference is blurred. If all possible observers do exist, the fact of my existence can’t be an evidence for the existence of a larger world.

Existence of all possible observers makes using of SSA more difficult, as I also have to account for my copies in the other worlds, which could also be the member of the reference class.

Combined, SSA and SIA in the multiverse tell us about the form of the distribution of the worlds: that it is quickly declining (that is Grace DA).

Such self-location uncertainty was used by Carrol in attempt to solve Born rule origins: “Quantum sleeping beauty” and “Self-Locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics.”

Metadoomsday argument for Sleeping Beauty

There several hundreds article about Sleeping Beauty problem – several times more than about DA. An agent with bounded rationality can’t expect that his own solution will outperform the solution given by most educated philosophers in our world. It is also obvious that collective human intelligence is not enough to solve the Sleeping Beauty problem in no doubts way.

Thus, he should give equal credence to halfers and thirders view. The first who mentioned that – according to my bounded knowledge – was Nancy Lebowitz: “I'm not sure whether this is legitimate or a joke, but if the question is unclear about whether 1/2 or 1/3 is better, maybe 5/12 is a good answer” in a comment to LessWrong post in 2009.

I also think that 5/12 is correct credence in such uncertainty situation. When I explained the SB problem to my 12 years old son, he also concluded this solution – may be inclination to solve the SB this way is genetically predisposed? Or I framed the SB problem in the way that such solution seems obvious.

Solving Sleeping Beauty problem anyway

First, we should mention that the world of SB is artificially constrained, for example, in the way that SB experiment can’t be repeat many times, and that there is no SB-copies in parallel world, but pure random coin anyway exists. It is not the way as our world is made, where there is a multiverse and quantum processes are the source of randomness.

In this world is postulated that SB could have credence in some events, but this credence is different from betting, as betting would make SB problem too simple. However, there is no practical difference in credence of 1/2 or 1/3 if betting can't be repeated.

In other words, SB is oversimplified thought experiment, based on contradicting ontology and it should not be surprising that it is spawning paradoxes.

Also, SB assumes that the probability of coins head is already known = 0.5, and in this point, it is different from the DA, where p of extinction is *a priori* unknown.

The main difference between SB and DA, is that in SB we know all setup and probabilities, but don't know our location, but in DA we don't know the setup, but know exactly our location and want to infer the setup from it.

To account for all this, we will modify the SB:

- There is random generator which output heads with unknown probability p, and tails with 1-p.
- If heads, 1000 copies of SB are created on Monday;
- if tails, 1 000 000 copies of SB will appear in each day after Monday.
- If SB is wrong about her location, she will be immediately punished by small amount of pain, so she is very interested to guess correctly, but she will not remember the result of guess in the next day.

In that setup, if SB don't know the day, she could conclude that she is most likely in the heads, if p is less than 0.99.

However, if she learns that it is Monday, when she should update her estimation of p to the highest level.

Appendix 3. DA objections analysis

1. Very large error on boundary conditions

DA gives more or less correct predictions for most observers, who are in the middle of the referent class, but it gives very large mistakes for those who are near the beginning, like those who were born on the 1st of January would expect that the year is only 2 days long.

2. Abnormal expectations as default model

If we take the Meadows model of the world (Meadows, Randers, and Meadows 2004)– that in the 21st century after exhaustion of natural resources, civilization and population will decline – there is nothing surprising in the DA. It just says that it is most probable to find oneself during the most populated period of the world history. This is similar to that fact that it is unlikely to be born in Lichtenstein, as it is very small country, and more likely the author was born in a large country.

Similarly, if we expect that one singleton AI will replace all humans during the Singularity – there is no problems with DA: it predicts exactly such a type of the end.

The problems with DA appears only in world model, where the world will be populated by billions human beings (or other human-like agents, like cyborgs with artificial consciousness) for millions of years, and will include also their space colonization. But this model is vanilla science fiction of the 20th century. Basically, DA predicts either a global catastrophe or an ascending superhuman AI which will transform (or kill) biggest part of the human population.

3. No new information from meta-DA

DA is a vague instrument for future predictions, and any practical data beats it. Meta-DA, which suggested that extinction soon is something like 25 per cent, produces the same order of magnitude as predictions about existential risks from such researchers as Bostrom and Leslie, which are generally shared by other researchers in the field. This is not surprising: the more actual data we get, the more likely our predictions (in most cases) will be the same as predicted by DA.

4. Order of the receiving of the information is important: it is surprising to find oneself first, but first one should not be surprised

Bostrom in his hybrid approach to the Sleeping Beauty problem demonstrated that the order of getting information is important for the probability estimates.

If someone learns first about the mediocrity principle and later learns about her/his unusual position in the set, it should surprise him, as it may mean a different size of the set. However, if someone learns first his unusual position, and later learns about mediocrity principle, s/he should not be surprised.

In our case, most people know from the childhood their date of birth and the fact that they are in the beginning of the possible human space history. Only after that some of them start to ask themselves, should they be surprised by this fact. And in that case, they should not.

However, this is not applicable to my position in the natural reference class, as I found this position only after I learned about DA.

Bostrom wrote about this problem of the order of getting information discussing the Sleeping Beauty problem (Nick Bostrom 2007).

Appendix 4. DA as a decision theory problem

If DA is true, so what?

In the group of theories called “Updateless decision theory” (also Timeless DT, Functional DT and Anthropic DT) the line of reasoning should escape updating on information of the local position of an individual, so all agents with the same line of reasoning will come to the same conclusion. This has positive consequences, as it produces cooperation between the agents in many decision theoretical problems, like Prisoners Dilemma [ref] and Newcomb problem [ref].

Functional decision theory suggests that one should treat oneself as random example of all calculation processes with the same setup (Yudkowsky and Soares 2017). One’s name, sex, location, etc generally do not typically affect the way one thinks about DA – they are random to the DA-style thinking.

So we should take all such “thought processes” and find the strategy where most of them win, without updating on the information about local position. This could be explained as the following: any observer who finds himself in the civilization, which is under global risk, should work on such risk prevention, and this will increase the total share of all civilizations which survive such risks.

Such observer thus should ignore her/his doubts about effectiveness of her/his own global risks prevention efforts, if such doubts arise from her/his special position in time.

However, updateless theories suffer from the problem of separation of “calculation process with the same setup” and random information about local position which should be ignored. For example, if I know DA – is it a random fact or part of the setup? The DA-aware reference class discussed above is a class specified around some knowledge about DA.

In the practical case of DA, one could use information implied by DA to update our behavior, so it will provide more effective survival: DA provides us with the information about the possible type and timing of a catastrophe. For example, universal DA tells us that universal catastrophic types should be taken seriously. Self-referenced DA implies timing of a catastrophe few decades from now.

Literature

Alasdair, M. Richmond. 2017. "Why Doomsday Arguments Are Better than Simulation Arguments."

Aranyosi, István A. 2004. "The Doomsday Simulation Argument. Or Why Isn't the End Nigh and You're Not Living in a Simulation."

Barrett, A.M., S.D. Baum, and K Hostetler. 2013. "Analyzing and Reducing the Risks of Inadvertent Nuclear War between the United States and Russia." *Science & Global Security* 21: 106–133.

Berezin, Alexander. 2018. "'First in, Last out' Solution to the Fermi Paradox." *ArXiv:1803.08425 [Physics]*, March. http://arxiv.org/abs/1803.08425.

Bostrom, N. 1997. "Investigations into the Doomsday Argument." *Preprint*, 359–387.

Bostrom, N. 2001. "The Doomsday Argument Adam & Eve, UN++, and Quantum Joe." *Synthese* 127 (3): 359–387.

Bostrom, N. 2002a. "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology, Vol. 9, No. 1 (2002).*

Bostrom, N. 2002b. "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation." *The Journal of Philosophy* 99 (12): 607–623.

Bostrom, N. 2003. "Are You Living In a Computer Simulation?" *Published in Philosophical Quarterly (2003) Vol. 53, No. 211, Pp. 243-255.*

Bostrom, N. 2007. "Sleeping Beauty and Self-Location: A Hybrid Model." *Synthese* 157 (1): 59–78.

Bostrom, N. 2013a. *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge.

Bostrom, N. 2013b. "Existential Risk Prevention as Global Priority." *Global Policy* 4: 15–31.

Bostrom, N. 2014. *Superintelligence*. Oxford: Oxford University Press.

Bostrom, N., and Milan M. Cirković. 2003. "The Doomsday Argument and the Self–Indication Assumption: Reply to Olum." *The Philosophical Quarterly* 53 (210): 83–91. doi:10.1111/1467-9213.00298.

Carroll, Sean M. 2017. "Why Boltzmann Brains Are Bad." *ArXiv:1702.00850 [Astro-Ph, Physics:Gr-Qc, Physics:Hep-Th, Physics:Physics]*, February. http://arxiv.org/abs/1702.00850.

Carter, Brandon. 1974. "Large Number Coincidences and the Anthropic Principle in Cosmology." In *Symposium-International Astronomical Union*, 63:291–298. Cambridge University Press.

Caves, Carlton M. 2000. "Predicting Future Duration from Present Age: A Critical Assessment." *Contemporary Physics* 41 (3): 143–153.

Chalmers, D. 1996. *The Conscious Mind*. Oxford University Press, New York.

Ćirković, Milan M., and V. Milošević-Zdjelar. 2003. "Extraterrestrial Intelligence and Doomsday: A Critical Assessment of the No-Outsider Requirement." *Serbian Astronomical Journal*, no. 166: 1–11.

Ćirković, Milan M., A. Sandberg, and N. Bostrom. 2010. "Anthropic Shadow: Observation Selection Effects and Human Extinction Risks." *Risk Analysis, Vol. 30, No. 10, 2010*.

Dieks, Dennis. 1992. "Doomsday–Or: The Dangers of Statistics." *The Philosophical Quarterly (1950-)* 42 (166): 78–84.

Egan, Andy. 2007. "Some Counterexamples to Causal Decision Theory." *The Philosophical Review* 116 (1): 93–114.

Garrabrant, Scott, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. 2016. "Logical Induction." *ArXiv Preprint ArXiv:1609.03543*.

Garrabrant, Scott, Tsvi Benson-Tilsen, Andrew Critch, Nate Soares, and Jessica Taylor. 2017. "A Formal Approach to the Problem of Logical Non-Omniscience." *ArXiv Preprint ArXiv:1707.08747*.

Gerig, Austin, Ken D. Olum, and Alexander Vilenkin. 2013. "Universal Doomsday: Analyzing Our Prospects for Survival." *Journal of Cosmology and Astroparticle Physics* 2013 (05): 013.

Google Scholar. 2018. "Google Scholar Data for Doomsday Argument." https://scholar.google.ru/scholar?start=70&hl=en&as_sdt=2005&sciodt=0,5&cites=2661658602441132437&scipsc=.

Gott III, J. Richard. 1993. "Implications of the Copernican Principle for Our Future Prospects." *Nature* 363: 315–319.

Gott, J. R. 1999. "The Copernican Principle and Human Survivability." *Human Survivability in the 21th Century. Transactions of the Royal Society of Canada, Series VI* 9: 131–147.

Grace, K. 2010. "SIA Doomsday: The Filter Is Ahead | Meteuphoric." *Meteuphoric*. https://meteuphoric.com/2010/03/23/sia-doomsday-the-filter-is-ahead/.

Grace, K. 2017. "When Will AI Exceed Human Performance? Evidence from AI Experts." https://arxiv.org/pdf/1705.08807.pdf.

Hanson, Robin. 1998. "Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization." https://philpapers.org/rec/HANBTC.

Kant, Immanuel. 1781. *Critique of Pure Reason*. Cambridge university press- 1999.

Knobe, Joshua, Ken D. Olum, and Alexander Vilenkin. 2006. "Philosophical Implications of Inflationary Cosmology." *The British Journal for the Philosophy of Science* 57 (1): 47–67.

Korb, Kevin B., and Jonathan J. Oliver. 1998. "A Refutation of the Doomsday Argument." *Mind* 107 (426): 403–410.

Korotayev, Andrey. 2018. "The 21st Century Singularity and Its Big History Implications: A Re-Analysis." *Journal of Big History* 2 (3): 73–119.

Kurzweil, Ray. 2006. *Singularity Is Near*. Viking.

Lem, Stanislaw. 1970. "Solaris. 1961." *Trans. Steve Cox and Joanna Kilmartin. New York: Hartcourt Brace*.

Leslie, John. 1996. *The End of the World: The Science and Ethics of Human Extinction*. Psychology Press.

Leslie, John. 2008. "Infinitely Long Afterlives and the Doomsday Argument." *Philosophy* 83 (4): 519–524. doi:10.1017/S0031819108000867.

Lewis, Peter J. 2013. "The Doomsday Argument and the Simulation Argument." *Synthese* 190 (18): 4009–4022.

Manheim, David, and Scott Garrabrant. 2018. "Categorizing Variants of Goodhart's Law." *ArXiv Preprint ArXiv:1803.04585*.

Meadows, Donella, Jorgen Randers, and and Dennis Meadows. 2004. *Limits to Growth: The 30-Year Update*. Chelsea Green Publishing.

Monton, Bradley. 2003. "The Doomsday Argument without Knowledge of Birth Rank." *The Philosophical Quarterly* 53 (210): 79–82.

Olum, K.D. 2002. “The Doomsday Argument and the Number of Possible Observers.” *Philosophical Quarterly* 52 (207): 164–184.

Pereira, Toby. 2017. “An Anthropic Argument against the Future Existence of Superintelligent Artificial Intelligence.” *ArXiv:1705.03078 [Cs]*, May. http://arxiv.org/abs/1705.03078.

Rees, M. 2003. *Our Final Century*. Heinemann.

Sandberg, A., Eric Drexler, and Toby Ord. 2017. “Dissolving the Fermi Paradox.” Future of Humanity Institute. http://www.jodrellbank.manchester.ac.uk/media/eps/jodrell-bank-centre-for-astrophysics/news-and-events/2017/uksrn-slides/Anders-Sandberg---Dissolving-Fermi-Paradox-UKSRN.pdf.

Simpson, Fergus. 2016. “Apocalypse Now? Reviving the Doomsday Argument.” *ArXiv Preprint ArXiv:1611.03072*.

Soares, Nate, and Benjamin A. Levinstein. 2017. *Cheating Death in Damascus*. Technical report, Machine Intelligence Research Institute, 2017. https://intelligence. org/files/DeathInDamascus. pdf. https://intelligence.org/files/DeathInDamascus.pdf.

Sober, Elliott. 2003. “An Empirical Critique of Two Versions of the Doomsday Argument–Gott’s Line and Leslie’s Wedge.” *Synthese* 135 (3): 415–430.

Sowers Jr, George F. 2002. “The Demise of the Doomsday Argument.” *Mind* 111 (441): 37–46.

Strugatsky, Arkady, and Boris Strugatsky. 1974. *Definitely Maybe*. Melville House, 2004.

Tainter, Joseph. 1990. *The Collapse of Complex Societies*. Cambridge university press.

Tegmark, Max, and N. Bostrom. 2005. *How Unlikely Is a Doomsday Catastrophe?* Vol. 438. 754. Nature, 1: 438-754. https://arxiv.org/abs/astro-ph/0512204.

Torres, P. 2016. “Agential Risks: A Comprehensive Introduction. 2016.” *Journal of Evolution and Technology* - 26 (2).

Turchin, A. 2018a. “‘Cheating Death in Damascus’ Solution to the Fermi Paradox.” *LessWrong*. https://www.lesswrong.com/posts/R9javXN9BN5nXWHZx/cheating-death-in-damascus-solution-to-the-fermi-paradox.

Turchin, A. 2018b. “The Risks Connected with Possibility of Finding Alien AI Code During SETI.” *Journal of British Interplanetary Society* 70.

Turchin, A. 2018c. “Forever and Again: Necessary Conditions for the ‘Quantum Immortality’ and Its Practical Implications.”

Turchin, A., and D. Denkenberger. 2018a. “Global Catastrophic and Existential Risks Scale.” *Futures, in Press*.

Turchin, A., and D. Denkenberger. 2018b. “Near-Term and Medium-Term Global Catastrophic Risks of the Artificial Intelligence.” *Artificial Intelligence Safety And Security*, *(Roman Yampolskiy, Ed.)*, *CRC Press*.

Turchin, A., B. Green, and D. Denkenberger. 2017. “Multiple Simultaneous Pandemics as Most Dangerous Global Catastrophic Risk Connected with Bioweapons and Synthetic Biology.” *Under Review in Health Security*.

Wallace, David. 2012. *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford University Press.

Ward, Peter D., and Donald Brownlee. 2003. *Rare Earth: Why Complex Life Is Uncommon in the Universe*. New York: Copernicus.

Weintraub, Ruth. 2009. “The Doomsday Argument Revisited.”

Wells, Willard. 2009. *Apocalypse When?: Calculating How Long the Human Race Will Survive*. Popular Science. Praxis. //www.springer.com/us/book/9780387098364.

Wilson, Alastair. 2015. “The Quantum Doomsday Argument.” *The British Journal for the Philosophy of Science* 68 (2): 597–615.

Yudkowsky, E. 2009. “The Anthropic Trilemma.” http://lesswrong.com/lw/19d/the_anthropic_trilemma/.

Yudkowsky, E. 2016. “Updateless Decision Theories.” Arbital. https://arbital.com/p/updateless_dt/.

Yudkowsky, Eliezer, and Nate Soares. 2017. “Functional Decision Theory: A New Theory of Instrumental Rationality.” *ArXiv:1710.05060 [Cs]*, October. http://arxiv.org/abs/1710.05060.

Zabell, Sandy L. 1989. “The Rule of Succession.” *Erkenntnis* 31 (2–3): 283–321.

Куркина, Е. С. 2013. “Конец Режимов с Обострением. Коллапс Цивилизации.” *Электронный Ресурс: Http://Spkurdyumov. Narod. Ru/Kurkinaes. Htm*.