A Pin and a Balloon:

Anthropic Fragility Increases Chances of

Runaway Global Warming

*Alexey Turchin,*

Foundation Science for Life Extension

[alexeiturchin@gmail.com](mailto:alexeiturchin@gmail.com)

*Anonymous coauthor 1*

*Anonymous coauthor 2*

**Abstract**: Humanity may underestimate the rate of natural global catastrophes because of the survival bias (“anthropic shadow”). But the resulting reduction of the Earth’s future habitability duration is not very large in most plausible cases (1-2 orders of magnitude) and thus it looks like we still have at least millions of years.

However, anthropic shadow implies *anthropic fragility*: we are more likely to live in a world where a sterilizing catastrophe is long overdue and could be triggered by unexpectedly small human actions. In the same way, an over-inflated toy balloon, which will soon burst, is very fragile.

Anthropic fragility can manifest itself in the higher chances of runaway global warming. It has often been suggested that the Earth's atmosphere remained life-supporting for billions of years by sheer chance. Therefore, the survival bias can be strong. It is also known that Earth-like water worlds could experience transitions into deadly moisture greenhouse (mean T = 65C). This means that relatively small anthropogenic actions could put the climate above an unpredictable tipping point, which could lead to the moisture greenhouse. Thus, it is necessary to carry out urgent geoengineering studies and prepare to prevent an unexpected climate catastrophe.

There are three main counterarguments against the existence of the anthropic shadow: self-indication assumption (SIA), past observers and the Gaia hypothesis; we show that they fail. It was proposed that SIA exactly compensates the anthropic shadow as an observer unlikely to find herself in a world with a strong anthropic shadow; however, there is a baseline level of the anthropic shadow for all habitable planets, similar to the rate of evolutionary transitions like abiogenesis. There are no “past observers” as qualified observers appeared only 50 years ago. Gaia hypothesis assumes existence of self-stabilizing feedback in climate, but new types of events like quick CO2 growth could override its coping ability.

We present a list of other catastrophes that may have been underestimated because of the anthropic shadow, including collider catastrophes, nuclear war and even an alien invasion.

We also hypothesized that human intelligence is more likely to emerge in an unstable world that is nearing its end and thus we get a new form of Doomsday argument.

**Highlights**:

· The decrease of the expected habitability time of the Earth’s because of the survival

bias (anthropic shadow) is not very large and thus it looks like that we still have at least millions of years.

· But anthropic shadow implies anthropic fragility: we are more likely to live in a world where a catastrophe is long overdue and could be triggered by unexpectedly small human actions.

· Catastrophic runaway global warming may be an example of such anthropic fragility, and we cannot predict where the tipping point is.

· We should research urgent geoengineering to counter the unpredictable anthropic fragility of climate.

· We also suggested a hypothesis that human intelligence is more likely to arise in an unstable world that is nearing its end.

[1. Introduction 3](#_Toc96087606)

[Previous literature review 4](#_Toc96087607)

[2. The central argument and a toy example 5](#_Toc96087608)

[2.1. Toy example 5](#_Toc96087609)

[2.2. The central argument 7](#_Toc96087610)

[3. Anthropic fragility and global catastrophes affected by it 7](#_Toc96087611)

[3.1. Anthropic fragility: underestimation of the fragility of our environment because of the anthropic shadow 7](#_Toc96087612)

[3.2. Runaway global warming as a most dangerous form of the anthropic fragility 8](#_Toc96087613)

[3.2.1. Moist greenhouse, tipping points and sea-floor methane 8](#_Toc96087614)

[3.2.2. Runaway Greenhouse 10](#_Toc96087615)

[3.2.3. CO2 concentrations which could start runaway warming 10](#_Toc96087616)

[3.2.4. Urgent geoengineering may be needed to counter climate models’ uncertainty 12](#_Toc96087617)

[3.3. Collider experiments as another type of anthropic fragility 12](#_Toc96087618)

[3.4. Different types of known global catastrophes and their anthropic shadows 14](#_Toc96087619)

[4. Counterarguments against the anthropic shadow 16](#_Toc96087620)

[4.1. Self-Indication Assumption counterargument: an observer is less likely to be in a world with high anthropic shadow 16](#_Toc96087621)

[4.2. “Early observers” counterargument 18](#_Toc96087622)

[4.2.1. Early observerhood 18](#_Toc96087623)

[4.2.2. Qualified observers 19](#_Toc96087624)

[4.2.3. Time until nuclear war 19](#_Toc96087625)

[4.3. Gaia counterargument 20](#_Toc96087626)

[4.4. PETM argument against runaway global warming 21](#_Toc96087627)

[5. Estimating the power of anthropic shadow on Earth 21](#_Toc96087628)

[5.1. Three types of power of anthropic shadow 21](#_Toc96087629)

[5.2. The similar size of the positive and negative fine-tuning 22](#_Toc96087630)

[5.3. Anthropic bias and the hard steps in the evolution 22](#_Toc96087631)

[5.4. How could one know about the existence of the anthropic shadow? 23](#_Toc96087632)

[6. Discussion: anthropic shadow and its connection with the evolution of intelligence 24](#_Toc96087633)

[6.1. Hypothesis: intelligence is more likely to appear in an unstable world that is close to its end 24](#_Toc96087634)

[6.2. Relation of the anthropic fragility to Fermi paradox and x-risks 25](#_Toc96087635)

[Conclusion 25](#_Toc96087636)

[Literature: 25](#_Toc96087637)

[Appendix 1. Different types of anthropic shadow 30](#_Toc96087638)

[Random events and regression to the mean 31](#_Toc96087639)

[Periodic catastrophes 31](#_Toc96087640)

[Semi-periodic catastrophes, which accumulate and then release energy 32](#_Toc96087641)

[Appendix 2*.* Normally distributed anthropic shadow 33](#_Toc96087642)

[Toy example with an old man 33](#_Toc96087643)

[Table 2. Life expectancy of survival if normally distributed catastrophe is long overdue. 34](#_Toc96087644)

[Box. The analysis of the toy example 35](#_Toc96087645)

[Mathematical model 38](#_Toc96087646)

[Life expectancy based on anthropic bias is approximately the same for different catastrophe distributions 40](#_Toc96087647)

[Appendix 3. Other ways to estimate the power of anthropic shadow 41](#_Toc96087648)

[5.4. Anthropic shadow magnitude could be estimated by frequency of near-misses 41](#_Toc96087649)

[5.5. Testing the “anthropic shadow” on the known catastrophes 42](#_Toc96087650)

[5.6. The estimation of the frequency of the natural catastrophes via the Fermi paradox 43](#_Toc96087651)

[Cuts: 44](#_Toc96087652)

[To add in new revision: 44](#_Toc96087653)

[How we can estimate the increase of probability of runaway global warming if we take into account anthropic shadow 44](#_Toc96087654)

[Tyrrell 45](#_Toc96087655)

[Example with car dealership 45](#_Toc96087656)

[Asteroid example and Gott 46](#_Toc96087657)

[2.4. Anthropic shadow has low sensitivity to the initial parameters 46](#_Toc96087658)

[3.2.2. Anthropogenic changes are different from natural 48](#_Toc96087659)

[6.1. Compression of evolution time in the case of a strong anthropic shadow 49](#_Toc96087660)

# 1. Introduction

One formulation of the anthropic principle (Carter, 1974) is that observers can exist only in those worlds in which there are no conditions that prevent the appearance of the observers. Such conditions may take the form of fine-tuning of the initial parameters of the universe – or the absence of life-ending catastrophes. Circovic et al called the lack of life-ending catastrophe in the past “anthropic shadow”, as it causes underestimation of the background rate of catastrophes.

The goal of this article is to research how anthropic shadow affects the fragility of our world, in the sense of Bostrom’s vulnerable world hypothesis (2018), especially relative to the risk of runaway global warming.

Our central argument and illustration are presented in section 2. We explore the idea of anthropic fragility and how it could work in different types of catastrophes, first of all, for global warming in section 3. We shortly discuss what types of geoengineering may be needed to counter unexpected rapid global warming. We will create an overview of all possible types of catastrophes where anthropic shadow and anthropic fragility can manifest themselves in table 1 in section 4. Then we will go into counterarguments that were suggested against the anthropic shadow and will demonstrate that they can’t completely disprove it, but they limit its power. We will explore what kind of evidence of anthropic shadow in the past of Earth we have in section 5. In the end, we will take more a general overview of the problem and of its connection with x-risks. We will suggest a new form of the Doomsday argument: that intelligent life is likelier to appear in the unstable world which is close to its end.

### Previous literature review

Bostrom and Tegmark wrote:

One might think that since life here on Earth has survived for nearly 4 Gy (Gigayears), such catastrophic events must be extremely rare. Unfortunately, such an argument is flawed, giving us a false sense of security. It fails to take into account the observation selection effect that precludes any observer from observing anything other than that their own species has survived up to the point where they make the observation. Even if the frequency of cosmic catastrophes were very high, we should still expect to find ourselves on a planet that had not yet been destroyed. The fact that we are still alive does not even seem to rule out the hypothesis that the average cosmic neighborhood is typically sterilized by vacuum decay, say, every 10,000 years, and that our own planet has just been extremely lucky up until now. If this hypothesis were true, future prospects would be bleak”. (Tegmark & Bostrom, 2005)

However, they also demonstrated based on the relatively late appearance of Earth in the history of the Universe that the space catastrophes capable of sterilizing Earth should be relatively rare, occurring no more often than once in 1 billion years with 99.9 per cent confidence. This is an example of SIA counterargument discussed in Section 4, which unfortunately does not work for Earth habitability.

Circovic, Sandberg and Bostrom wrote a subsequent article, “Anthropic Shadow: Observation Selection Effects and Human Extinction Risks” in 2010 (Ćirković, Sandberg, & Bostrom, 2010), in which they created a Bayesian update equation to calculate the actual probability of a natural global catastrophe which takes into account observation selection effects. In that article, they listed five different types of natural catastrophes which observed probabilities could be affected by observation selection: asteroid/comet impacts, solar superflares, supervolcanic eruption, close flyby of rogue black holes and nearby supernovae explosions or gamma-ray bursts.

However, Ord et al (Snyder-Beattie et al., 2019) suggested that the natural catastrophes background rate for Homo Sapiens is not affected by the observation selection effects, as there could be *earlier observers* (this will be discussed in section 4.2).

Manheim applied anthropic bias analysis to the frequency of natural pandemics in the past (Manheim, 2018).

Meanwhile, A. Scherbakov noted that the history of the Earth’s atmosphere is strangely correlated with the solar luminosity and the history of life, which could be best explained by anthropic fine-tuning, in the article “Anthropic principle in cosmology and geology” (Shcherbakov, 1999). In particular, he wrote that the atmospheric temperature was closely preserved in the range of 10–40 °C, and on four occasions the Earth came close to a “snowball” steady-state, and on four occasions came close to turning into a water vapor greenhouse where the temperature could reach of hundreds of degrees centigrade. However, these life-ending outcomes were prevented by last-minute events such as volcanic eruptions or covering of volcanoes in the ocean by water, which regulates the CO2 level following an eruption. Such “miracles” are best explained by observation selection effects.

Waltham looked at the Milankovitch cycles and using the modelling of a group of planets found that the Solar system has unusually low orbit perturbations: “…the probability of all three occurring by chance is less than 10−5. It therefore appears that there has been anthropic selection for slow Milankovitch cycles. This implies possible selection for a stable climate, which, if true, undermines the Gaia hypothesis and also suggests that planets with Earth-like levels of biodiversity are likely to be very rare” (Waltham, 2011).

In the article “On the absence of solar evolution‐driven warming through the Phanerozoic” Waltham argues in favor of anthropic explanations of climate stability: “The Gaia hypothesis, anthropic selection or some other unconventional mechanism may therefore have to be invoked to explain the absence of long‐term warming through the Phanerozoic” (Waltham, 2014).

In 2020, two important articles on the topic were published. The article by Tyrrell “Chance played a role in determining whether Earth stayed habitable” (Tyrrell, 2020) demonstrated via computer modelling of the history of many Earth-like planets that chance played a significant role in preserving the stability of the atmosphere.

Another article, “The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare” (Snyder-Beattie et al., 2020) is analyzing the timing of the important steps of the evolution of life and suggests that the observed frequency of such steps has strong anthropic effects, and median frequency is likely much lower, which implies longer expected time of development of intelligent life; from this follows Rare Earth. Robin Hanson in the “Grabby aliens” article discussed similar ideas (Hanson et al., 2021).

Wordsworth (Wordsworth, 2021) showed that runaway global cooling (snowball Earth) is still probable and could start if the global temperature falls to 7 C, and it was just a few degrees more than that only 20 000 ago at the peak of the Ice Age. He regards it as a possible explanation of the Fermi paradox, as in the snowball Earth glaciers will reach the equator and will destroy all complex life.

Survival bias is clearly ignored in some estimations of the risks of runaway global warming:

“The paleoclimate record can be used to check our prediction that surface temperatures might increase dramatically were they to exceed ~305 K (Fig. 4b). The closest analog to the climate regime modeled here is the mid-Cretaceous Period, ~100 million years ago, which is considered to be the warmest Earth has ever been during the Phanerozoic eon. Life clearly existed and flourished during this time, so we know that whatever surface temperature prevailed then was not high enough to trigger a moist greenhouse.” (Ramirez et al., 2013)

# 2. The central argument and a toy example

## 2.1. Toy example

Imagine that every 100 million years *on average* a sterilizing catastrophe is happening on any Earth-like planet. It is a random process with a half-life of 100 of millions years. It is completely unobservable if it didn’t happen.

We suggest here as a toy example that such a catastrophe is a nuclear explosion of a natural nuclear reactor in the Earth core, which some hypothesized to exist (Herndon, 1993). The explosion of the reactor will not destroy the planet but it will produce strong shockwaves and volcanism which will result in complete surface replacement in Earth, like the one which has happened in Venus. As Earth has existed for 4.5 billion years, the chances of the Earth to survive until now is 1 in 245 =3.5x1013. That is, anthropic shadow on Earth with the power around 1013 corresponds to the future life expectancy of 100 million years.

If the reactor explodes with the half-life of 10 million years, the power of anthropic shadow is 1 in 10135, which is truly immense. But for evolutionary transitions like abiogenesis, such probabilities are not uncommon and we will show later (section 6) that the power of anthropic shadow is likely to be similar to the power of evolutionary transitions. Totani estimated the chances of abiogenesis and calculated that only 1 of 10100 planets will generate self-replicating RNA via randomness (Totani, 2020).

Even if the reactor explodes every once in one billion years, we have one billion years of future life expectancy, but anthropic power is 1 in 16. If there will be no anthropic shadow then expected future survival will be equal to the past survival and will be 4.5 billion years.

It seems that human civilization will be fine, as we still will have 10 million years (with 50 per cent probability) in the worst case, which is enough to become a multi-planetary civilization.

However, any small additional fuel will trigger the explosion. Russian scientists suggested sending a probe to the Earth’s core based on a small nuclear reactor that melts all the way down (Ozhovan et al., 2005). If the natural reactor is “trigger-happy”, such experiment could be enough to destabilize it. Such an Earth-core probe will be like a pin that is poking an overinflated balloon – dangerous game!

Remind you, that it is a toy example, so the real probe unlikely will reach the core, but it is still could cause large a degasation event (Cirkovic & Cathcart, 2003).

Moreover, the problem is that the nuclear reactor inside Earth could have accumulated the fuel all that time. We suggested before that its explosion is a truly random effect, but for some processes like inflating balloons or overstretched springs, the probability and the power of explosions grow non-linearly in time. This means a significantly shorter future life expectancy and higher fragility. Here we looked at *exponentially distributed* risks. We look at *normally distributed* catastrophes in Appendix 2 and we assume there that for an overinflated-balloon-type of catastrophe the probability is *normally* distributed around some mean value. In that case, for example, 14 sigmas anthropic shadow event has 1 in 1045 chances. It produces the future habitability of Earth equal to 0.04 per cent of the past age, but it is still 1.6 million years half-life which provides enough time for a civilization to leave Earth.

If there are two universes, and in one of them all Earth-like planets have half-life 100 million years because of the core reactor explosions, and in another universe, it is 200 million years, – at the end, there will be 222.5=500 million timesmore habitable planets in the second universe. Therefore, we are more likely to find ourselves in the universe with a weaker anthropic shadow. The second universe provides the baseline level of the anthropic shadow if only two types of universes exist. This is the *SIA counterargument* in a nutshell which will be discussed later in section 4.

This thought experiment demonstrated to us the following properties of the anthropic shadow:

*1. Even a minor anthropic shadow means a significant cut of future life expectancy*.

*2. The sensitivity of the future life expectancy to the power of anthropic shadow is relatively small. The growth of anthropic shadow for 10134* times lowers the life expectancy by only 100 times.

*3. Any anthropic shadow means significant growth of fragility.*

*4. The power of the anthropic shadow is limited to some baseline* which applies to all Earth-like planets.

5*. Non-linear probability distributions of catastrophes mean much shorter future life expectancy time and higher fragility*.

Humanity will either become an interplanetary civilization in the next millennia, or it will never do it, given the expected quick growth of nanotech and AI. Therefore, a million years timescale of natural catastrophes is not itself a significant risk for our space future.

## 2.2. The central argument

The toy example above helps us to formulate the central argument about the anthropic fragility of the habitability of the Earth, which we will explore later in detail. The argument runs according to the following lines:

1. Habitable planets with intelligent life in the Universe are rare because many types of catastrophes could kill life.
2. The fact that we are alive means that we were very lucky and have escaped many past catastrophes.
3. But therefore, we can’t estimate the frequency of such sterilizing catastrophes based on the observations. This is the anthropic shadow.
4. If some catastrophe is long overdue, this lowers our future life expectancy for around 1-2 order of magnitude, based on the low sensitivity of anthropic shadow to initial parameters. This seems to be not problematic as it still gives us millions of years of life expectancy.
5. However, the fragility of our environment also has grown, and thus relatively small or *unnatural* anthropogenic actions could cause an unexpected catastrophe.
6. The primary risk here is a sudden climate catastrophe caused by crossing an unexpected tipping point which will start a positive feedback loop and will increase Earth’s temperature to 65 C moisture greenhouse level or even higher. Collider accidents and nuclear war are other examples of anthropic fragility.
7. To counter this *unobservable fragility*, we need to be more careful and to have quick reaction instruments, like urgent geoengineering.
8. However, as intelligence is a universal adaptation, it is more likely to evolve in the world with changing climate, and such changes themselves are the sign that the climate catastrophe is near. This increases the chances of catastrophe soon. More on that in section 6.

# 3. Anthropic fragility and global catastrophes affected by it

## 3.1. Anthropic fragility: underestimation of the fragility of our environment because of the anthropic shadow

Now it’s time to look deeply at what is anthropic fragility. Not all types of survival bias result in anthropic fragility. If a plane with holes from an enemy fire has returned to the base, there is no anthropic fragility. Thus, anthropic fragility requires that the situation of risk is *continuous*. But even this is not enough: asteroid impact risk is continuous, but there is no anthropic fragility, as there are no meaningful ways how humans could affect the probability of large impacts, except asteroid deflection. Poking an overinflated balloon is an example of the fragility, and injuries from falling for an older man is another ­– but all these examples are not anthropic: there is no observation selection effect which results in the underestimation of risks.

If a plane with many holes is still flying back to its base, it is closer to anthropic fragility, as the plane is damaged and any sharp turn could put too much force on its structures and will fail apart.

Anthropic fragility is the following situation:

1. *Anthropic shadow.* There is a strong anthropic shadow, so we live in a world where some kind of otherwise very probable sterilizing catastrophe has not yet happened by pure chance.
2. *Parameter*. The probability of the catastrophe is connected with a slow accumulation of some parameter, similar to the pressure in an inflating balloon.
3. *Tipping point*. The catastrophe will inevitably happen if the parameter will achieve some threshold level (or at least, catastrophe’s probability significantly increases if the parameter grows even slightly).
4. *Human actions*, like experiments or emissions, are changing the value of the parameter, so it could reach the threshold. Human actions may affect some other foundations of stability that don’t look fragile at first glance. These indirect human actions must be *unique* in the sense that they never happened before in exactly the same way.
5. *Unknown to humans*. Humans do not know the real activation level of the parameter because they can’t infer its value from the past rate of catastrophes, which they never observed because of the survivorship bias. They may even not know that such a catastrophe is possible at all.

Anthropic fragility is defined both as *physical* and *epistemic* situations: the physical situation is that a catastrophe is long overdue, and the epistemic part is that we can’t know it from the past observation and therefore underestimate the safe level of the changes of the parameter. Even if we agree with the idea of anthropic fragility, we still can’t know which parameter is causing fragility and what are its safe levels of manipulation. In the damaged plane example, the pilots may know that the plane is fragile, but do not know which actions are risky: climbing, turning or landing.

There are two types of anthropic fragility: one is when we increase the pressure on something which is already under pressure, that is, like adding more air in the inflated balloon, and another is when we perform unique, never happened before actions on a system which is in a metastable condition, like poking the balloon with a pin. We could call them a “parameter increase” and “unique actions”. In the case of global warming, adding more CO2 and increasing temperature is the “parameter increase”; and the unprecedented speed of warming and unique methane effects are an example of unique human actions. In reality, both types of fragility are connected, because unique actions cause some parameter increase or the parameter increases in some unique way.

For example, if our false vacuum is very close to the transition to the next metastable state, Large Hadron Collider (LHC) experiments may be more dangerous than they appear (Hut & Rees, 1983) because they create some unique type events, as was suggested by Kent, more on that below.

## 3.2. Runaway global warming as a most dangerous form of the anthropic fragility

### 3.2.1. Moist greenhouse, tipping points and sea-floor methane

A recent article suggested, based on computer simulation, that water-world planets (which Earth is similar to) have a second semi-stable temperature regime, *moist greenhouse,* with a mean temperature of 57°C while Earth now has a mean temperature of 15°C (Popp et al., 2016). All temperature regimes between these two will collapse to either the current climate or to a moist greenhouse; thus, there should be some *tipping point* between these two temperatures after which positive feedback loops dramatically accelerate.

We assume here that the moist greenhouse will cause human extinction, the same way as ocean evaporation could do it, so there is no practical difference between the two from anthropic and existential risks views. However, there is a small possibility to survive moist greenhouse in some coldest places on Earth like very high mountains, like Himalayas, and in the Antarctic. Toby Ord estimated in the *Precipice* (Ord, 2020) that there is a 0.1 per cent probability of the existential risk this century because of climate change, mainly due to the start of the moist greenhouse.

The idea of tipping points in climate has been often discussed, but because of anthropic shadow, the position of the actual location of the tipping point could be underestimated. In a climate context, “tipping point” means not only a temperature but some combination of temperature, greenhouse gases concentrations, albedo, solar luminosity and sea-floor methane release rate. The limited size of the previous “close-call” tipping points, like PETM warming, could be explained by the observation selection effect: if the Earth had turned into a moist greenhouse 55 million years ago, there would be no observers now.

The question of the speed and the possibility of methane-driven climate change is a topic of scientific debate (Shakhova et al., 2010) which are too complex to be completely presented here. But anthropic shadow effect of climate fragility increases our uncertainty about such feedback loops. If the transition takes only a few years, we will not have time to cope with it via geoengineering, except perhaps via a nuclear explosion in a supervolcano to create artificial volcanic winter. However, the anthropic shadow prevents us from observing quick and large magnitude changes of the climate in the past, if they are possible, so the absence of such events in our history is not evidence for further slow future climate changes.

The anthropic fragility becomes especially important relative to the catastrophes which *are long overdue*. An overinflated toy balloon is in a metastable state, where even a small punch could lead to its explosion. Higher levels of radiation from the Sun has been compensated by historically low levels of CO2, which, however, helped glaciation and methane accumulation in the permafrost and ocean floor, which later could cause a very large “clathrate gun” rapid warming (Kennett et al., 2003). This fragility can’t be observed in the historical record because of anthropic shadow. We should be extremely careful with climate change.

The main difference between now and previous periods of warming is the large accumulation of methane hydrates which by some estimates are 10 times higher now than during PETM, as current ice age conditions helped such accumulation on the seafloor and in the permafrost (Ananthaswamy, 2015, 2015; Dean et al., 2018). The topic of risks from methane eruption (Ananthaswamy, 2015; Dean et al., 2018) from the Arctic is controversial: some claim that it is the main risk of warming and other present models that methane leaking will not be enough to start runaway global warming. Anthropic shadow could make us underestimate the power of previous methane discharges. Also, there is a difference between anthropogenic and natural effects on methane: as methane is a short-lived gas in the atmosphere, its concentration depends on the speed of its leaking from reservoirs, which itself depends on the speed of the temperature change. Anthropogenic global warming is relatively quick, because of the unprecedented speed of CO2 emissions and thus will produce more methane concentrations than the same CO2-driven warming if it were slower. Recent research showed that the speed of change was a characteristic of past mass extinctions (Song et al., 2021).

Many argue that climate change is not an existential risk and that its danger is exaggerated (Lomborg, 2020). But what makes it a real existential risk is the fat tail of uncertainty of its magnitude powered by anthropic bias.

### 3.2.2. Runaway Greenhouse

Besides moisture greenhouse, which is hypothetically survivable, there is an even worse scenario, when whole oceans evaporate and equilibrium temperature reaches 1400K, which is described in the article “The Runaway Greenhouse: implications for future climate change, geoengineering and planetary atmospheres” (Goldblatt & Watson, 2012):

The ultimate climate emergency is a “runaway greenhouse”: a hot and water vapor rich atmosphere limits the emission of thermal radiation to space, causing runaway warming. Warming ceases only once the surface reaches ∼1400 K and emits radiation in the near-infrared, where water is not a good greenhouse gas. This would evaporate the entire ocean and exterminate all planetary life. Venus experienced a runaway greenhouse in the past, and we expect that Earth will in around 2 billion years as solar luminosity increases.

Goldblatt & Watson concluded that CO2 emissions alone can’t cause this, but if other warming sources will add up, this becomes possible:

The question here is simply how much could human action increase the strength of the greenhouse effect? Kasting & Ackerman (1986) found that, with carbon dioxide as the only non-condensible greenhouse gas, over 10,000 ppmv would be needed to induce a moist greenhouse. This is likely higher than could be achieved by burning all the “conventional” fossil fuel reserves—though the actual amount of fossil fuel available is poorly constrained, especially when one includes “exotic” sources such as tar sands (which are already being exploited). Greenhouse gases other than carbon dioxide, cloud or albedo changes could all contribute further warming. Likewise, the exhibition of multiple equilibria in the relevant temperature range (Renn´o, 1997; Pujol & North, 2002) complicates matters.

They conclude: We cannot therefore completely rule out the possibility that human actions might cause a transition, if not to full runaway, then at least to a much warmer climate state than the present one.

High climate sensitivity might provide a warning.” (Goldblatt & Watson, 2012) and then argue that we may need geoengineering to stop runaway warming. Growing climate sensitivity could be a warning sign. However, they validate the sensitivity estimates on past climate without taking into account a possible anthropic shadow: “Such high sensitivity is inconsistent with our knowledge of paleoclimate and the model cases which provide the extremes do not seem likely (due to poor representation of contemporary climate)”.

Therefore, the question is: are existing climate models are capable to predict risks of runaway warming? If yes, we should not worry about the anthropic fragility of climate. But the models are limited:

Ideally, we would want numerical climate models to robustly resolve the transition to a much hotter atmosphere. However, most such models have been developed for fairly small perturbations from the existing climate and their wider applicability may be limited by the obvious unavailability of data to tune the model to, and by simplifications made to reduce computational cost… A new generation of model may well be needed (Collins et al., 2006; Goldblatt et al., 2009)).

### 3.2.3. CO2 concentrations which could start runaway warming

We could measure *anthropic fragility* as an amount of deviation from the current level of some parameter, which will cause a global catastrophe and which estimation is distorted by the anthropic shadow. For example, how much CO2 could be added to the atmosphere before a *tipping point* is reached?

The Earth in the past had CO2 levels much higher than today, so a *naïve* view is that even having 10 times more CO2 than now will not cause runaway global warming. But this view does not take into account the anthropic effects, that is, we cannot use the past data about the probability of runaway global warming, as we could observe ourselves only on a planet where runaway global warming never happened. Also, previous higher levels of CO2 were at least partially compensated by lower Sun’s luminosity, lower deposits of methane hydrates and other geophysical factors then, such as the different configurations of oceans and different parameters of Earth orbit. Such factors in the past worked as protection against runaway global warming, but they are not in place now.

There are different assessments of the critical levels of CO2 after which runaway warming will happen. As was cited above 10 000 ppm is needed for a moisture greenhouse. Another estimation of the dangerous level is 30 000 ppm CO2 which is unachievable by anthropogenic mineral fuel consumption (Goldblatt et al., 2013), and another estimation is only 12 times the preindustrial level, that is, 3360 ppm (Ramirez et al., 2014). The current CO2 level seems to be the highest in the last 5 million years ago and could become higher than it was in the whole Miocene (20 million years long) in the next 100 years (Dean et al., 2018).

According to IPCC’s, in the worst-case SPP5-8.5 scenario the CO2 levels could reach 1000 ppm at the end of the 21 century (Allan et al., 2021), which seems to be not enough to trigger runaway global warming, according to cited above scenarios.

However, the real problem of the anthropic shadow is that we can’t know for sure that this CO2 level will not cause runaway warming. There are several sources of uncertainty:

* Large uncertainty comes from the possibility of methane emissions from the permafrost and other sources. Methane warming potential and release rate depend on *the speed of warming,* as methane is short-living gas: if the release rate is low, methane concentration will be also low. Human anthropogenic warming is different from the natural increase of CO2 in the past as anthropogenic CO2 levels are increasing quicker, and thus the role of methane will be higher.
* Other pollutants, like N2O, also have high warming potential.
* Our efforts to cut emissions can backfire, because of aerosol-removal effects, as industrial aerosols block part of sunlight in the upper atmosphere
* More generally speaking, our models are calibrated on the historical relation between CO2 and global temperature, but because of the anthropic shadow, this relation could be unreliable. We survived only in those worlds where extreme deviations never happened.
* Chaotic nature of climate. In the same way as weather, climate also could be fundamentally unpredictable. There could be strong deviations from the mean and bifurcation between different semi-stable attractors.
* Unpredictable combination of random events may force Earth out from the zone of climate stability. One such event was Younger Dryas when a large ice lake leaked.
* Our models may be incomplete in long run and may not take into account some factors.

We are not trying to say that the IPCC model is wrong. It is the best what we have. What we try to show is that there is a small probability of a *tail risk*, which, however, has most of the expected negative value.

Also, the nature of the tipping point is that it is *the beginning of a self-accelerating runaway process*, but not the point where the process will reach its maximum level. Thus, the CO2-concentration when the runaway warming starts is not the same as the one at which the CO2 level will be high enough to cause tens degrees of warming and human extinction. Methane and water vapor may be the main drivers of warming after reaching the tipping point, so the lack of CO2 sources will not stop the warming. This is similar to poking of a small hole by a needle in a toy balloon, which will inevitably result in its explosion. The needle creates a hole but does not participate in the process of destruction after that.

Fragility can be illustrated by an example of overstretched spring, in which fragility and life expectancy are measured in the same units: per cents of additional inflation. Therefore, the anthropic fragility should be measured not in the CO2 levels, but in *the temperature increase*, as CO2 levels are non-linearly proportional to warming and depends on other things that could affect temperature. There were [periods](https://www.science.org/content/article/500-million-year-survey-earths-climate-reveals-dire-warning-humanity?fbclid=IwAR0S9OUFLvCQDqeP68ecfKAZrgfatg46L1FxTGJ71iXaNoFOA03eibTFMLo) when the Earth temperature was 15C higher than now and the hypothetical moisture greenhouse temperature is 45C higher than current temperatures, so the naïve view is that we are safe even if maximum expected global warming will happen with around 5-8C temperature increase. Thus, the tipping point for moisture greenhouse is somewhere between 15C and 45 C increase, according to the naïve view.

However, the idea of anthropic fragility means that we should lower our estimates for around an order of magnitude, which, given all uncertainties, means that the tipping point could lie not in tens but in *single digits* of temperature increase (that is, between 1.5C and 4.5C, if we just divide on 10 the above estimate). This temperature increase could be reached in the 21st century, and maybe even in the next decade. See the discussion below why PETM warming 55 million years ago is not a safe data point, as at that time was different conditions than now.

How long it takes from the tipping point to full-blown moisture greenhouse is unclear, and it could be from weeks, if water vapor feedback is activated, to centuries, if methane and ocean thermal inertia will play a major role (Karnaukhov, 2001). It is also unclear, if the warming will stop on the moisture greenhouse level, or will overshoot and will go to the Venusian runaway scenario.

### 3.2.4. Urgent geoengineering may be needed to counter climate models’ uncertainty

As humans do not—and cannot because of the anthropic shadow effects—know the location of the climate tipping points (Lenton, 2011) based on the past historical record, it may be prudent for society to perform more research about the climate tipping points and to research methods of *urgent geoengineering*, which could give us more time if runaway warming starts. One model gives 4 years estimate until moisture greenhouse from the moment of the beginning of radiative forcing (Seeley & Wordsworth, 2021).

The only such measure which could be used on short notice is the use of nuclear weapons to start nuclear winter by causing large fires in taiga or by the initiation of volcanic eruptions. Such measure could put us back below the tipping point, or give a few more years to prepare better protection measures. Higher altitude explosions over taiga will have less nuclear fallout and could be done using existing delivery systems. The slower method is the use of existing airplanes to perform sulfate stratospheric injection (Halstead, 2018). “Normal” geoengineering aimed at the gradual removal of CO2 does not give us enough flexibility to react to unexpected climate changes. We also should do everything else possible to prevent anthropogenic global warming, first of all, cutting emissions.

## 3.3. Collider experiments as another type of anthropic fragility

Another type of fragility is associated with the risk of conducting physical experiments that create completely new conditions on Earth, i.e. experiments at the hadron colliders. It has been suggested that they could cause three types of catastrophes: a false vacuum collapse that would end the entire observable universe, a mini-black hole that would slowly but acceleratingly eat Earth's matter, and a special type of quark matter called strangelets, which supposedly capable of turning ordinary matter into strange matter (Kent, 2004).

One popular argument for collider safety is based on the observed safety of high-energy collisions of cosmic rays with Earth’s atmosphere, but it could also suffer from anthropic effects. The argument is that if experiments at colliders could create mini-black holes, then cosmic rays would be capable of it. However, since we have survived for billions of years, this has never happened. Therefore, high-energy collisions cannot create anything dangerous. But this argument clearly suffers from the fact that it does not take into account the survival bias.

Dar et al. (1999) presented an "anthropically invulnerable" argument for the safety of the collider: if high-energy collisions are dangerous, we could observe random supernovae from sudden collapses of other stellar objects and planets. Note that Dar's argument works only for the formation of strangelets and small black holes, which will only cause a local catastrophe on Earth, but not for the collapse of a false vacuum that can destroy the entire universe.

In the case of collider catastrophes, anthropic fragility plays a role, since human experiments turn out to be different from natural phenomena and, thus, can cause a catastrophe not observed on other planets. A. Kent wrote that there is a subtle difference between collider experiments with collisions of cosmic rays with the Earth's atmosphere. Namely, the products of collider collisions have zero velocity relative to the Earth, because they are produced by two opposite particle beams, and the products of cosmic ray collisions continue to move at near-light velocity relative to the Earth. This may change the nature of the interaction of products with terrestrial matter (Kent, 2004) giving mini-black holes more time to accumulate mass and start growing.

The main question here is: are the experiments at the collider *unique* or not in the sense that these are events that have never happened in the history of the Earth? If they are not unique and similar events have occurred on Earth and other planets without causing the end of the world, there is no risk of anthropic fragility.

Bostrom and Tegmark explored the probability of the *natural false vacuum decay* and similar cosmological catastrophes and found it to be low given our late existence. Their article (Tegmark & Bostrom, 2005) does not directly apply to the collider experiments, as it puts a limit only on the *natural* cosmological catastrophes, and if the colliders create really *unique* conditions, no natural process is a reference class; of course, it also means that there are no alien civilizations in our past light cone which had performed such dangerous experiments before us and had created a wave of destruction through the whole universe. Their counter-arguments are discussed in more detail in section 4.1.

In 2008, it was suggested that the series of collider failures that occurred before the launch of the LHC could be explained by the anthropic shadow, since we can only survive on worlds where the collider does not work; see also (Ord et al., 2010) about anthropic considerations in the collider risk estimations. However, LHC started its operations after that and continued for around 10 years. However, the LHC began its work after that and lasted for about 10 years. If the LHC had a high probability of causing a false vacuum decay, then our own existence as the authors of this article seems too late, based on the same line of argument used by Bostrom and Tegmark: if LHC causes vacuum breakdown, most scientists should find themselves just after the beginning of its operation (or before). This is also an example of a more general counter-argument against the anthropic shadow called "early observers", which is discussed below in section 4.2.

One way or another, the anthropic shadow can make us underestimate the fragility of our vacuum, and any unique (which has never been in nature) experiment can become a pin that can burst the “over-inflated ball” of an over-due catastrophe.

## 3.4. Different types of known global catastrophes and their anthropic shadows

Waltham (Waltham, 2019) explored seven possible anthropic fine-tunings of the Solar System, including the mass of the Sun, the mass of the Moon, the orbits of the major planets, the Earth's ocean, the Earth's magnetism, and the Earth's plate tectonics. The evidence for the fine tunings he found is not very strong when taken individually, but collectively they point to some anthropic pressure in the evolution of the solar system. However, most of the fine-tuning methods he explores are positive conditions; here we are interested in negative conditions, that is, what types of events should not occur for the emergence of intelligent life. We have included in Table 1 several other global risks which have been hypothesized to be affected by the anthropic shadow.

*Table 1. Types of possible catastrophes and corresponding anthropic effects*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Type of**  **natural catastrophe** | **Description and comments** | **Evidence for anthropic pressure** | **Could observed catastrophes rate be affected by anthropic shadow?** | **Could it have anthropic fragility?** | **Type of period for periodic events** |
| 1. Collider catastrophe   (Ord et al., 2010; Sandberg, 2008) | Colliders hypothetically could create new types of matter which could destroy Earth: mini black holes, strangelets or false vacuum transition (Kent, 2004) | Sandberg suggested that the LHC's series of failures in the early 2010s is best explained by the "quantum immortality effect" because if it starts working, it will destroy the world. | Yes | Yes, in the form of uniqueness of human experiments | No |
| 2. False vacuum decay (Hut & Rees, 1983) | Hypothetical type of catastrophe | String theory suggests 10500 possible vacuums.  (Douglas, 2003); Many of them are unstable (Adler et al., 1995) | Yes | Hadron collider may trigger such a catastrophe | Not known; probably random event |
| 3. “Alien invasion” (Gertz, 2016), (Hanson et al., 2021), (Turchin & Denkenberger, 2019) | The shock wave of alien colonization, at a speed close to the speed of light, consumes all matter in the universe. (Armstrong & Sandberg, 2013). | We have only been able to survive in that part of the universe where aliens have not yet appeared or are hiding.  Aliens may be closer than Fermi's Paradox suggests. | Yes | Messaging to Extra Terrestrial Intelligence could attract hostile aliens if they are nearby  (Baum et al., 2011) or we could find their dangerous signals (Turchin, 2018c). | If natural intergalactic panspermia (spread of life) is possible, civilizations of our age are closer in space and could arrive on Earth relatively soon. (Turchin, 2020). |
| 4. Gamma-ray burst (GRB)  (Ćirković & Vukotić, 2016) | Directed GRB could sterilize planets 1000 light years away. | The sun is located in a relatively distant part of the galaxy, and terrestrial life arose relatively late in the history of the universe, so gamma-ray bursts are rare in our region. | No, because we are seeing a lot of distant gamma-ray bursts. However, some nearby stars may be on the verge of forming a burst that was "delayed" by the anthropic shadow. | No | Random event |
| 5. Nearby supernova (Ćirković & Vukotić, 2016) | Supernova is dangerous if it is closer than 8 light years (Ćirković & Vukotić, 2016). | We are not in that part of the Galaxy where there are many supernova explosions (in the core). | No, as we observe many remote supernovas. | None | Random, very rare event based on observational constraints |
| 6. Solar super-flare  (Lingam & Loeb, 2017) | Sun-like stars have super-flares which could be dangerous to life on Earth. | The Sun seems to be a surprisingly quiet star, as other stars of its type are more active  (Reinhold et al., 2020). Also, Sun may have had larger flares in the past but now is quiet. | If the Sun were to become more active, higher levels or radiation and electromagnetic storms might prevent the rise of a technological civilization. | None | Not known |
| 7. Asteroid/comet impact (Chapman & Morrison, 1994) |  | Over the past few hundred million years, there have been no collisions that could have wiped out all vertebrates. (Rampino & Caldeira, 2015) | There is a periodicity of about 30 million years of mass extinctions that could be caused by an influx of comets from the Oort cloud, and we are now living near the end of such a period.  In addition, several stars have recently passed close to the Sun and may have disturbed the Oort cloud. (Bailer-Jones, 2015). | None | Random cometary streams could be semi-periodic if they were associated with perturbations of the Oort cloud by solar oscillations around the galactic plane. |
| 8. Supervolcanic eruptions and flood-basalt events  (Rampino, 2008). | Large-scale eruptions, i.e. Siberian traps; even larger eruptions are possible, resulting in surface replacement | There have been no comparable basalt flood events since the Great Dying 242 million years ago, except perhaps for Deccan traps. |  | Could be triggered by experiments with deep drilling and core penetration but currently unlikely. (Cirkovic & Cathcart, 2003), | Tension-spring mode |
| 9. Runaway global warming  (Popp et al., 2016). |  | The peculiar stability of the Earth's climate, despite the change in the luminosity of the Sun. |  | May be caused by increased CO2 and methane emissions from permafrost. | Inflated balloon mode |
| 10. Ocean anoxic event and H2S poisoning of atmosphere  (Ward, 2007). |  | Black sea could produce enough H2S to poison atmosphere and the level of H2S in it is growing  (Kump et al., 2005) |  | Phosphorus is the main problem. It is used as a fertilizer throughout the world and ends up in the oceans. This could trigger an anoxic event in the ocean, which has been associated with mass extinction events, probably due to the formation of H2S (Handoh, 2013). Global warming increases the likelihood of anoxic events. |  |
| 11.Stellar encounters with giant molecular clouds  (Kokaia & Davies, 2019). | This can lead to a closer supernova, an increase in the frequency of asteroid impacts, and climate change. | Such encounters affect the stability of the Oort cloud and the planet's climate through dust accretion. | Anthropic effects define the galactic habitable zone where such encounters are rare. | No | Random, 1.6 event in Gyr for Sun. |
| 12.Nuclear war  (Sandberg et al., 2018). | Some have suggested that the fact that World War III did not occur in the 20th century is best explained by anthropic selection. | A nuclear war is unlikely to kill everyone, so the anthropic effects should not be strong. | A nuclear war would most likely destroy civilization and there would be fewer scientists left to discuss anthropics, so we would most likely be discussing it in a world where there was no World War III. | The risk of an accidental nuclear war and the ease of provocation can be underestimated because of survival bias. | No. |
| 13.Next Ice age | Human civilization was able to appear only during a period of warm and stable climate which was good for agriculture and large empires (Gowdy, 2020). It has been relatively warm and stable for the last 11 700 years after the Young Dryas. It was predicted that the next Ice age will happen in a few millennia from now. The previous interglacial period’s duration was around 13 ky and happened 130 ky ago. |  | Human civilization arose during a stable interglacial period, and such periods, as a rule, make up only one tenth of the total time of the modern glaciation. | Humanity's efforts to combat global warming could unexpectedly backfire and lead to an early onset of the next cooling period. This includes CO2 reduction and geoengineering. | Periodic event with a random component |

# 4. Counterarguments against the anthropic shadow

## 4.1. Self-Indication Assumption counterargument: an observer is less likely to be in a world with high anthropic shadow

The counterargument states that if there are two possible worlds, and one of them has an anthropic shadow and the other does not, then there will be more observers in a world without an anthropic shadow, and so I am more likely to be in a world without such a shadow.

In a nutshell: *The anthropic shadow is exactly compensated by my smaller chance to be in such world.*

Thus, according to this counterargument, we are likely to live in a world where there is no survival bias and no problem with underestimating the risks of future disasters.

But there are three objections to this counterargument which are limiting its power: minimal level of anthropic shadow for all worlds; the numerical dominance of the semi-fine-tuned worlds and a possible correlation between catastrophes and habitability which will be discussed below.

But firstly, a counterargument’s example: imagine that there are two groups of potentially habitable planets, and in each group initially there were 100 planets. In the first group, the probability of the past catastrophes is 0, and in the second group, only 1 planet has survived and gave rise to a civilization. Thus, in the second group, the anthropic pressure is 100 to 1. However, if an observer does not know in which world she is located, she has 100 times greater chances to be in the first group of worlds, as there will be 100 habitable planets.

This reasoning is based on so-called Self-Indication Assumption (SIA) which favors the worlds with larger number of observers (Bostrom, 2013). The SIA counterargument was suggested by Stuart Armstrong in a comment to my blog post about anthropic shadow.

SIA have two interpretations: first, it is simply about the distribution of observers among different universes in the multiverse, but it requires that all these universes actually exist. In another, stronger interpretation, the SIA is seen as an argument that of two possible universes, the more populated one is more likely to actually exist like in the Presumptions philosopher thought experiment (Bostrom & Ćirković, 2003). In case of modal realism, these two versions merge. Here we will use the first, simpler one, which assumes the existence of many universes with different properties. For more on SIA and observer densities see in my article “Presumptuous philosopher proves panspermia” (Turchin, 2020).

In other words, it looks like that SIA exactly compensates the anthropic shadow, the same way as it exactly compensates the estimation of future catastrophes following from the Doomsday argument (Bostrom & Ćirković, 2003).

For example, if in the one part of the multiverse the false vacuum decay is likely and in the another it is not possible, the observer is much more likely to find herself in the part of multiverse where false vacuum decay never happens.

We can generalize this principle:

*The anthropic shadow’s power is limited by the possibility of existence of another universe with less anthropic shadow which produces the same observer.*

However, this principle cannot exclude the anthropic shadow idea completely, as some anthropic shadows may be *irreducible*.

For example, the anthropic shadow has limited the size of the asteroids which have impacted the Earth in the last billion years, but the existence of asteroids and impacts may be a necessary condition for the evolution of life on Earth-like planets, as a large impact likely created the Moon (Barr, 2016), comets may have brought water to Earth (Hartogh et al., 2011), and the impact-driven extinction of dinosaurs was probably necessary for mammals to rise to dominance (Alvarez et al., 1980). Thus, all possible habitable planets are likely under the risk of large impacts. Because of this, intelligent life may have evolved only on a small share of lucky planets; thus, the anthropic shadow for impacts would be strong and future expected rate of impacts would be higher.

Therefore, the first objection to the SIA-counterargument is that there is *a minimal level of anthropic shadow for all habitable planets*.

Another objection to SIA counterargument is the idea that the worlds with anthropic shadow could still produce more planets with civilizations than the worlds without such shadow as they are more numerous. To illustrate this, we will return to our example above. Before, we assumed that the initial number of the planets was equal. Now let us assume that initially there are 10 000 planets with 0.01 survival rate, and 10 planets with 1 survival rate. At the end, there will be 100 habitable planets in the first group, and still 10 in the second. In that case, we are still more likely to find ourselves on a planet that had a strong anthropic shadow. Let us call the first group *semi-fine-tuned* for intelligent life, and the second one – *fine-tuned*.

In that case, *the* *anthropic shadow is real if the share of semi-fine-tuned planets in the multiverse is larger than the share of the fine-tuned ones*. Obviously, there are no data about such shares, but mathematical simulation of possible planets by Tyrrell gives similar results: most currently habitable planets will have anthropic shadow (Tyrrell, 2020).

The observational consequence of this is that most observers may appear not because of the perfect fine-tuning of initial conditions, but because of some very random events which overcome non-perfect fine-tuning. In a fictional example, in a perfectly fine-tuned universe there would be no dangerous asteroids/comets in stellar systems. But in our solar system, we have many asteroids, and this could be compensated by some additional conditions, such as Jupiter protecting Earth from comets, or by pure chance. If humanity is protected by pure chance, this protection may not work in the future, which means a higher chance of asteroid/comet impacts in the future.

In other words, the domination of imperfectly fine-tuned worlds means that negative fine-tuning—where the random absence of the catastrophic events is the main mechanism of civilizational survival—may be the dominating form of anthropic selection, and thus the anthropic shadow would be much stronger.

The third reason to reject the SIA-counteragent is that there could be *a correlation between higher risk of catastrophes and habitability*. Universes which allow *interstellar panspermia* will have billion times more planets with life, but in such universes asteroid impacts are more frequent and stars are closer to each other, which implies higher rate of natural catastrophes (Turchin, 2020). Another idea of this type is that intelligence is more likely to appear in an unstable world which will discussed in section 6.

However, *SIA is a strong argument against extreme forms of anthropic bias*, like the collapse of false vacuum every 10 000 years, but doesn’t work against weaker forms of anthropic bias, like higher rate of impacts or a possibility of runaway global warming. It is not easy to say without concrete data where is the line of balance between SIA and anthropic bias.

We think that SIA allows some form of anthropic bias, maybe as weak as 1 order of magnitude, but as we showed in section 2, the magnitude of anthropic bias has a relatively small impact on the future life expectancy, which is around 0.1 of previous time, if some form of anthropic shadow is present. However, in section 5, we will explore an argument that minimal level of anthropic shadow is comparable with probability of other big evolutionary filters and is many orders of magnitude.

## 4.2. “Early observers” counterargument

### 4.2.1. Early observerhood

In the article “An upper bound for the background rate of human extinction” Snyder-Beattie et al. (Snyder-Beattie et al., 2019) introduced the idea of *early observerhood* which could appear long before now: “To model observation selection bias, let us assume that after Homo sapiens first arises another step must be reached. This could represent the origin of language, writing, science, or any relevant factor that would transition early humans into the reference class of those capable of making observations (we call this step ‘observerhood’).” They assume that such a factor may appear something like 20 000 years ago, while Homo Sapiens has existed for 200 000 years.

After long calculations, they conclude: “In summary, observer selection effects are unlikely to introduce major bias to our track record of survival as long as we allow for the possibility of early observers”. The main reason for this is that if observers have been here for a long time, it is a strong argument against the high background extinction rate, as otherwise, one would likely find oneself near the beginning of the existence of observers. For example, if observers have been on Earth for 20 000 years, this is a strong counterargument against the background extinction rate of 1 in 1000 years, as only 1 in 220 timelines will reach such stage.

As one should think about oneself as an observer randomly selected from all observers, one should find oneself where most observers are concentrated, that is, in the first millennium in this example. But humanity does not find itself at this time, according to Snyder-Beattie’s logic, so 1 in 1000 extinction rate per year is most likely wrong.

In a nutshell, it is a Doomsday argument in reverse, as we use observers’ long survival as an argument against the high rate of catastrophes (Bostrom, 2001; Turchin, 2018a).

### 4.2.2. Qualified observers

However, it looks like that Snyder-Beattie et al. used a too relaxed definition of a “qualified observers” that constitute the relevant reference class. We argue that the qualified observer should be able to understand the idea of the “anthropic shadow” and be able to think about related topics. Neither hunter-gatherers nor medieval writers could do this. People with the necessary mathematical training and similar ideas began to appear only in the 18th and 19th centuries, i.e. Laplace with his sunrise problem (Marquis De Laplace, 1814), but the number of such observers increased only at the end of the 20th century when anthropic reasoning and Doomsday argument appeared.

The reason we use a more rigorous definition of observerhood is the idea that “I am randomly selected among functionally indistinguishable observers”, that is, from all those who think about a certain topic (Yudkowsky & Soares, 2017). We should ignore in our calculations all observers who *are not thinking about this topic*, no matter if they are bright minds, can speak, can feel or whatever.

Therefore, there are no “early observers” (with one exception discussed below). Currently living population of the observers is the first one, and because of that, they are especially unsure about their past and future extinction rates, like Adam and Eve in Bostrom’s article (Bostrom, 2001). Moreover, it now looks like an argument in favor of higher extinction risks: as qualified observers exist for only around 50 years, future life expectancy could be also short.

Despite our criticism of the definition of observerhood, Snyder-Beattie’s counterargument applies only to the natural risks that could have occurred in the last 20 000 years, where the difference between a current person’s location and earlier observers is significant. It can’t be applied to long-term risks, such as the rate of asteroid impacts or stability of the atmosphere, with reoccurrence rate of dangerous situations around tens of millions of years. Snyder-Beattie’s counterargument also doesn’t work for climate’s anthropic fragility, as anthropogenic global warming has started only in last couple of hundred years.

### 4.2.3. Time until nuclear war

However, there is one situation where Snyder-Beattie’s counterargument works. It is anthropic estimate of the probability of nuclear war. There was no global nuclear war for more than 75 years, and anthropic reasoning exists from the 1970s, as Carter suggested anthropic principle in 1973, that is 48 ago from now (as of 2022).

If nuclear war has median timing of around 10 years, we are unlikely to find ourselves so late. Every-100-years-nuclear war hypothesis seems to be more likely in such a situation than every 10 years hypothesis. Every-1000-years nuclear war again does not look probable because now the question arises: why we are so yearly? In other words, if we have 3 hypotheses about the typical frequency of nuclear war: 10, 100, or 1000 years from the creation nuclear weapons and assume that there are no other x-risks, then finding ourselves 75 years after the creation of nuclear weapons supports the 100-years hypothesis.

Even if nuclear war is not killing all the people, it will significantly reduce the number of qualified observers, as the biggest university centers will be affected and Internet will not appear. Therefore, nuclear war could be regarded as equal to the extinction of qualified observers.

Bostrom’s argument against the frequent false vacuum decay is based on the relatively late time of Earth formation in the history of our galaxy (Tegmark & Bostrom, 2005). It is also an example of the “early observerhood” counterargument. Its logic can’t be applied to the catastrophes on Earth, as we don’t have evidence that life on Earth can evolve quicker than it did; moreover, we have opposite observations about the frequency of evolutionary transitions on Earth.

## 4.3. Gaia counterargument

The hypothesis of *Gaia* (Lovelock & Lovelock, 2000) suggests that climate can self-regulate via negative feedback loops and that earth life is an important part of this self-regulation. An example of such feedback loop may be the growth of the Earth’s albedo because of the larger amount of clouds if temperature grows too much. Life plays important role in the self-regulation of climate as it is able to capture CO2.

The appearance of Gaia also could be explained by anthropic selection (Tyrrell, 2020; Watson, 2004), as only planets capable of self-regulating climate have preserved their habitability. Gaia cannot be explained via the Darwinian selection on Earth as was suggested by (Doolittle, 2019), as Gaia exists only in one example. The planets with life that did not evolve a homeostatic mechanism of climate regulation would not exist for long and therefore would not be capable of producing observers.

Gaia seems to protect against anthropic fragility of climate, as most of the anthropic selection for stability has already happened in the past and has produced a self-stabilizing mechanism, now independent of anthropics. However, it may be that Gaia is just an observational selection illusion: there is no self-regulation, but there is just a series of random events which helped our survival.

Generalizing the Gaia counterargument, one may say: *all anthropic selection has happened in the past and this has selected a system that is very stable and has homeostatic mechanisms supporting its stability to a wide range of perturbations*. The theory may claim that homeostatic systems are rare, but eventually, they become more numerous than the worlds which just randomly escape all possible dangers. E.g. Jupiter is like Gaya for asteroid collisions, as it cleans space from hazardous comets.

Tyrrell demonstrated that only some planets get Gaya-like mechanism of homeostasis, and such mechanisms still have limits, above which these planets go into runaway warming (Tyrrell, 2020).

The main objection to this counterargument is that any self-regulation has its limits. The Sun will eventually overheat the Earth, which is typically estimated to happen around 1 billion years from now if one does not account for anthropic effects and feedback climate loops, which could imply earlier runaway global warming.

Biological organisms also age, become fragile and die and Gaia can age too. Thus, Gaia also may be unable to cope with some unexpected blows when “planetary boundaries” are exceeded (Baum & Handoh, 2014) and may stop protecting us from climate change.

Unique anthropogenic actions could unexpectedly end Gaia, as its protection works only for natural variations. Human-caused deforestation and other interventions in nature could affect Gaia coping ability and increase the probability of greenhouse catastrophe as was discussed in “Role of the Biosphere in the Formation of the Earth’s Climate: The Greenhouse Catastrophe” by (Karnaukhov, 2001).

## 4.4. PETM argument against runaway global warming

The PETM episode of global warming, when 55 million years ago global temperatures jumped 8 C (or even 15C according to recent research), probably because of methane eruption, could be presented as a counterargument to the danger of runaway global warming, as temperatures returned to normal.

However, the main idea of this article is that one cannot use evidence from the past as arguments for our future survival because of anthropic bias. Maybe the PETM had a 99 per cent chance to turn into runaway global warming, but one cannot observe this, so we observe only the timeline (or Everett branch) where this did not happen. Also, the situation during PETM was different when now: different amounts of methane deposits, different speeds of warming and different disposition of continents.

During the last interglacial period 130 ky ago, the temperatures were also warmer than now by 2C and this did not trigger methane-driven runaway global warming. But as we discussed above, the methane-driven warming will be significant only if the speed of the initial CO2-driven warming is high, because of the short lifetime of methane in the atmosphere. So, the condition then was different than now, and thus it is not proof that we are safe now.

# 5. Estimating the power of anthropic shadow on Earth

## 5.1. Three types of power of anthropic shadow

The power of anthropic shadow – that is, the chances of our survival until now – is not very important, as very different anthropic shadows give only 1-2 orders of magnitude reduction of future life expectancies, as we show in section 2. Such difference doesn’t strongly affect our decision-making. As we assume that fragility increase is proportional to the life expectancy decrease, it means that fragility variation is not very large.

Based on the power of anthropic shadow, we could distinguish 3 significantly different situations:

* anthropic shadow (ASH) doesn’t exist at all.
* *ASH is weak,* from one to few orders of magnitude. Future life expectancy decrease is around one order of magnitude and is not important for us in case of most natural catastrophes. Anthropic fragility is present but manageable. This type of ASH is favored by SIA counterargument. For example, 1 in 1010 chances of past catastrophes give 100 million years of future life expectancy instead of 1 billion years. 10 times increase in fragility means that, say, not 45C is needed to reach moisture greenhouse, but only 4.5C, which is still manageable by emission control.
* ASH is *strong*, many orders of magnitude. The decline of life expectancy of the biosphere is significant. The world is so fragile to human actions that the fatal damage is likely already happened. This type of ASH is favored by evolutionary transitions argument discussed below. For example, 1 in 10100 ASH gives only around 12 million years of life expectancy and two order of magnitude anthropic fragility. This means that only around 0.5C of anthropogenic temperature increases is enough for the start of runaway global warming, and we already past that point.

## 5.2. The similar size of the positive and negative fine-tuning

A recent article “The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare” by Sandberg et al (2020) shows that given a known future life expectancy for Earth's climate of 1 billion years, all evolutionary transitions could have a very small probability, or, in other words, very long mean times, maybe many orders of magnitude longer than the age of the universe. This means a very strong anthropic selection effect, since we are observing a world in which all such events occurred in time.

We suggest that *strong anthropic selection for evolutionary transitions means that the that an equally strong anthropic effect existed in avoiding natural disasters*. Although this is not easy to prove, we can illustrate this with the concept of "budget: if a person has an expensive house, he is likely also has an expensive car, since the expensive house suggests that the person has a large budget.

Similarly, powerful anthropic effects in evolutionary transitions may suggest the existence of a large "anthropic budget", that is, the actual existence of a large number of different worlds from which our world was chosen at random. The reason for this is that the Earth's longer survival time allows more time for evolutionary transitions.

Therefore, if it is easier to “buy” more time for the stable existence of a planet than to accelerate evolutionary transitions, then anthropic selection will choose planets with a longer existence without catastrophes. This will result in situation in which gaining time becomes as difficult as accelerating the evolutionary transition. Thus, there is a trade-off between avoiding catastrophes by pure chance and the speed of evolutionary transitions. For example, if the Earth were habitable not for 5, but for 10 billion years, this would allow more time for transitions, but would require much more incredibly long survival without catastrophes. The details of this trade-off are beyond the scope of this article and merit further study.

The similarity between anthropic shadow and evolution transitions probabilities is not exact, and it is not even of the same order of magnitude. It could be 1 in 10200 chances of intelligent life per planet and 1 in 10180 that a planet survives intact all risks of sterilizing catastrophes, like large asteroid impacts and runaway warmings. The second probability is trillions of trillions times higher, but still, they are *similar* in some sense.

In addition, the evaluation of fine-tuning by Sandberg et al. (2020) refer to *global* *base rate*, i.e. it is a probability distribution for all terrestrial planets in the Universe, not just for the Earth. If there is a region of the universe where evolutionary transitions are much easier, we will be there. Therefore, the estimation of the risk’s frequencies is also a base rate applicable to all planets, and thus SIA counterargument (section 4.1) can’t kill it. This means that everywhere in the universe there is a strong base rate of sterilizing catastrophes.

## 5.3. Anthropic bias and the hard steps in the evolution

As we showed above, the power of anthropic shadow is comparable with the difficulty of evolutionary transitions, and one more estimation of this difficulty is presented in the article by Kipping.

The article “An objective Bayesian analysis of life’s early start and our late arrival” (Kipping, 2020) compared the probability of abiogenesis, which took around 100-300 million years, and the arrival of intelligence, which required 3.8 billion years after the appearance of life and concluded that intelligence appearing was a more difficult and less probable event. But the abiogenesis itself is estimated to be a very improbable event: a recent estimate of the minimum length of self-replicating RNA is around 100 bases (Totani, 2020). Totani thinks that only one of 10^100 stars will generate a viable RNA strand. Thus, only one of 10^80 of Hubble volumes of the universe would have life. But intelligence would be even less probable, based on Kipping’s logic.

If intelligence is a rare step, it could be explained if the typical habitability of planets is shorter than the one of Earth and thus Earth is long overdue to lose its habitability. Thus, the world’s end is nigh, though we still may have around 100 million years. But anthropic fragility means this situation is worse as any significant change will cause a catastrophe.

However, this estimate may be inaccurate due to natural interstellar panspermia (Ginsburg et al., 2018), since in the case of panspermia, life may take much longer to develop. Life may have evolved on other planets before coming to Earth. For anthropic reasons, universes with panspermia will create more observers, as an entire galaxy with billions of potentially habitable planets(Hsu et al., 2019) could be fertilized with life (Turchin, 2020).

## 5.4. How could one know about the existence of the anthropic shadow?

Anthropic shadow may be hinted by several types of evidence, but its main feature is that it is mostly unobservable. These types of evidence are:

* *Frequency of near-misses.* One indication is the frequency of near-misses, or situations of improbable survival in the past. There is currently unpublished work “Nuclear war near misses and anthropic shadows” on the topic by Sandberg et al (Sandberg et al., 2018). For climate, it is several past Snow Ball situations, and one potential runaway greenhouse event 55 mln years ago (PETM), which didn’t turn permanent.
* *Absence* *of the fat tail.* Another indication of anthropic shadow is the absence of a fat tail in the distribution of smaller catastrophes. For example, we observed asteroid impacts of bodies only below some size; there was no 100 km bodies impacts in last couple of billion years on Earth. We also never observed that the Earth becomes too hot, like 60 C, as it seems to be irreversible tipping point for runaway moisture greenhouse.
* *Surprising coincidences.* A further evidence is some surprising coincidences which helped survival and evolution of life on Earth, as was analyzed in Sandberg et al (Snyder-Beattie et al., 2020). One of such evidence is the ability of Earth to have habitable temperature despite the growth of Sun luminosity.
* *Observation that we are located near the end of several period catastrophic cycles.* This is a situation where one finds oneself closer to the end of stability period, or even some signs of the end of stable period could be observed. For example, asteroid impacts’ waves seem to be semi-periodic events every 30 or so million years, probably cause by disturbance of Oort cloud from passing of Galactic plane.
* *Increasing instability.* Smaller catastrophes could be a sign that a larger one is near, e.g. increase of smaller asteroid impacts.
* *Mathematical modeling* of the similar to the Earth planets with different initial conditions (Tyrrell, 2020). This could provide *a priori* probability of anthropic shadow.
* General considerations about Rare Earth and Fermi paradox.

We will continue discussion about the different types of evidence for anthropics shadow in Appendix 3.

6. Discussion: anthropic shadow and its connection with the evolution of intelligence

## 6.1. Hypothesis: intelligence is more likely to appear in an unstable world that is close to its end

Above we discussed a weak version of the anthropic shadow. But in the same way, as the *strong anthropic principle* claims that the universe *needs* to create observers (Barrow & Tipler, 1986), we could formulate a *strong anthropic shadow*:

*Intelligence tends to appear only in a world that is close to its end*.

The main reason for the emergence of intelligence near the end of the world is that intelligence is a general adaptation that outperforms specialized adaptations in a rapidly changing world, which in the case of Earth is a world with an unstable climate. During *Homo sapiens’* evolution, humans’ predecessors changed several main ways of feeding in just a few million years (Henry, 2018). Our hominid ancestors started out as arboreal primates, then evolved into savannah scavengers, then into foragers, then—maybe—“aquatic apes” then fire-cooking hunters, then plants-eating early agricultural inhabitants. communities, and then again to the high-calorie eaters of modern cities. Periodic glaciations and deglaciations in the last few million years have affected human living conditions and create the need to adapt to new feeding opportunities.

The second reason that the end is near is that the evolution of intelligent observers requires unusually long periods of relative stability: thus, a very large catastrophe may be overdue. This may seem to contradict what we said above, but here we are referring to much larger catastrophes that can kill all humans or irreversibly destroy civilization, such as moisture greenhouse.

Intelligent observers (scientists) are much more fragile than the entire biosphere and even the species *Homo sapiens*. A large economic collapse could reduce their numbers, e.g. as it happened after the collapse of the USSR. This means, as Circovic et al. wrote, that anthropic shadow must be stronger in recent times and suppress even smaller catastrophes (Ćirković et al., 2010). This includes smaller asteroid impacts, supervolcanic eruptions and nuclear war.

The third reason is that the impact of human civilization on nature is growing exponentially. If there is some hidden vulnerability (Bostrom, 2018) that can provoke a global catastrophe, humanity will stumble upon it sooner or later, especially since the anthropic shadow does not let us know where such a vulnerability is. The anthropic shadow also results in underestimation of the fragility of our environment to small anthropogenic changes.

Waltham came to similar conclusions:

Given this link between climate change and species diversity, it is plausible that planets with high climate variability may be less likely to produce intelligent observers than planets with more stable conditions. However, it is also arguable that the ultimate emergence of intelligent species is actually encouraged by adverse conditions because these help to clear ecological niches (cf. the adaptive radiation of mammals following demise of the dinosaurs) and because evolutionary innovations may be particularly advantageous during testing times [cf. the emergence of *Homo sapiens* during the relatively unstable Neogene … and the emergence of multicellular life around the time of the Neoproterozoic glaciation] (Waltham, 2011).

In addition, human actions are unique and different from events that have occurred naturally in the past, and the homeostatic mechanisms that exist in nature may not be adapted to the new type of change. This includes the speed of change in the case of global warming and possibly a unique combination of several anthropogenic factors.

Another argument for a strong anthropic shadow is the Doomsday argument, which suggests a high level of future unknown risks based on our early location in human history (Bostrom, 2012; Turchin, 2018), and the general abundance of global catastrophic risks in our world (Bostrom, 2002).

In a sense, the anthropic shadow is similar to the idea of so-called *quantum immortality* where a person observes oneself surviving many rounds of Russian roulette because of the many world interpretation of quantum mechanics (Turchin, 2018) if this idea is applied to humanity’s past. Even if humanity is in a simulation, most simulations likely model a civilization near its end, and will terminate after modeling the “singularity” (Greene, 2018; Turchin, Yampolskiy, Denkenberger, & Batin, 2019).

Similar idea was suggested by global warming research pioneer Budyko who [discovered](https://alev-biz.livejournal.com/5333134.html) that cephalization of animals grew in the periods of unstable climate of Ice ages when forest and savanna replace each other with periodicity of tens of thousands of years and universal adaptive ability – intelligence – was in demand.

## 6.2. Relation of the anthropic fragility to Fermi paradox and x-risks

A strong anthropic shadow means that we are alone in the universe, so there is less risk of encountering hostile aliens. And vice versa, the observation that we alone, hints at a higher frequency of natural catastrophes that we have never seen due to the anthropic shadow.

Therefore, there is an increase of catastrophic risks in both branches of possibility: if we are alone, then we are exposed to higher natural risks and to anthropic fragility. If we are not alone, then the risk of hostile aliens is higher (Turchin, 2018c; Turchin & Denkenberger, 2019).

Anthropic fragility puts natural risks into the short-term perspective of the next 100 years. Therefore, they become comparable with the most serious technological risks: synthetic biology, AI and nuclear war.

# Conclusion

The anthropic shadow may seem like a purely theoretical effect, as far greater global risks pose threats to human civilization, such as unaligned AI, nuclear war, and synthetic biology. (Turchin & Denkenberger, 2018).

However, the detailed analysis presented here shows that anthropic shadow means that there are natural catastrophes that may be long overdue, have accumulated destructive energy, and could be triggered by small human actions. Such human actions will be unique events capable to trigger a change in semi-stable conditions. The most dangerous of these catastrophes is the runaway global warming that could occur if some currently unknown threshold of warming levels is reached.

# Literature:

Adler, R. J., Casey, B., & Jacob, O. C. (1995). Vacuum catastrophe: An elementary exposition of the cosmological constant problem. *American Journal of Physics*, *63*, 620–626. https://doi.org/10.1119/1.17850

Allan, R. P., Hawkins, E., Bellouin, N., & Collins, B. (2021). *IPCC, 2021: Summary for Policymakers*.

Alvarez, L. W., Alvarez, W., Asaro, F., & Michel, H. V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science*, *208*(4448), 1095–1108.

Ananthaswamy, A. (2015). The methane apocalypse. *New Scientist*, *226*(3022), 38–41. https://doi.org/10.1016/S0262-4079(15)30420-6

Armstrong, S., & Sandberg, A. (2013). Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. *Acta Astronautica*, *89*, 1–13.

Bailer-Jones, C. a. L. (2015). Close encounters of the stellar kind. *Astronomy & Astrophysics*, *575*, A35. https://doi.org/10.1051/0004-6361/201425221

Barr, A. C. (2016). On the origin of Earth’s Moon. *Journal of Geophysical Research: Planets*, *121*(9), 1573–1601. https://doi.org/10.1002/2016JE005098

Barrow, J. D., & Tipler, F. J. (1986). *The Anthropic Cosmological Principle (Clarendon*. Oxford.

Baum, S. D., & Handoh, I. C. (2014). Integrating the planetary boundaries and global catastrophic risk paradigms. *Ecological Economics*, *107*, 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024

Baum, S. D., Haqq-Misra, J. D., & Domagal-Goldman, S. D. (2011). Would contact with extraterrestrials benefit or harm humanity? A scenario analysis. *Acta Astronautica*, *68*(11–12), 2114–2129.

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

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

Bostrom, N. (2012). A Primer on the Doomsday Argument. *Anthropic-Principle.Com*. /?q=anthropic\_principle/doomsday\_argument

Bostrom, N. (2013). *Anthropic bias: Observation selection effects in science and philosophy*. Routledge.

Bostrom, N. (2018). *The Vulnerable World Hypothesis*. 38.

Bostrom, N., & Ćirković, M. M. (2003). The doomsday argument and the self-indication assumption: Reply to Olum. *The Philosophical Quarterly*, *53*(210), 83–91.

Bowen, G. J., Maibauer, B. J., Kraus, M. J., Röhl, U., Westerhold, T., Steimke, A., Gingerich, P. D., Wing, S. L., & Clyde, W. C. (2015). Two massive, rapid releases of carbon during the onset of the Palaeocene–Eocene thermal maximum. *Nature Geoscience*, *8*(1), 44.

Bridges, W. (2012). Gains from Getting Near Misses Reported. *Presentation at 8th Global Congress on Process Safety, Houston TX April*, 1–4.

Carpenter, P. A., & Bishop, P. C. (2009). A review of previous mass extinctions and historic catastrophic events. *Futures*, *41*(10), 676–682. https://doi.org/10.1016/j.futures.2009.07.012

Carter, B. (1974). Large number coincidences and the anthropic principle in cosmology. *Symposium-International Astronomical Union*, *63*, 291–298.

Chapman, C. R., & Morrison, D. (1994). Impacts on the Earth by asteroids and comets: Assessing the hazard. *Nature*, *367*(6458), 33–40.

Cirkovic, M. M., & Cathcart, R. (2003). Geo-engineering Gone Awry: A New Partial Solution of Fermi’s Paradox. *ArXiv Preprint Physics/0308058*.

Ćirković, M. M., Sandberg, A., & Bostrom, N. (2010). Anthropic shadow: Observation selection effects and human extinction risks. *Risk Analysis, Vol. 30, No. 10, 2010*.

Ćirković, M. M., & Vukotić, B. (2016). Long-term prospects: Mitigation of supernova and gamma-ray burst threat to intelligent beings. *Acta Astronautica*, *129*, 438–446. https://doi.org/10.1016/j.actaastro.2016.10.005

Cui, Y., Schubert, B. A., & Jahren, A. H. (2020). A 23 m.y. Record of low atmospheric CO2. *Geology*. https://doi.org/10.1130/G47681.1

Dar, A., De Rújula, A., & Heinz, U. (1999). Will relativistic heavy-ion colliders destroy our planet? *Physics Letters B*, *470*(1–4), 142–148. https://doi.org/10.1016/S0370-2693(99)01307-6

Dean, J. F., Middelburg, J. J., Röckmann, T., Aerts, R., Blauw, L. G., Egger, M., Jetten, M. S., Jong, A. E., Meisel, O. H., & Rasigraf, O. (2018). Methane feedbacks to the global climate system in a warmer world. *Reviews of Geophysics*.

Doolittle, W. F. (2019). Making Evolutionary Sense of Gaia. *Trends in Ecology & Evolution*, *34*(10), 889–894. https://doi.org/10.1016/j.tree.2019.05.001

Douglas, M. R. (2003). The statistics of string/M theory vacua. *Journal of High Energy Physics*, *2003*(05), 046–046. https://doi.org/10.1088/1126-6708/2003/05/046

Firestone, R. B., West, A., Kennett, J. P., Becker, L., Bunch, T. E., Revay, Z. S., Schultz, P. H., Belgya, T., Kennett, D. J., & Erlandson, J. M. (2007). Evidence for an extraterrestrial impact 12,900 years ago that contributed to the megafaunal extinctions and the Younger Dryas cooling. *Proceedings of the National Academy of Sciences*, *104*(41), 16016–16021.

Flis, A. (2020, October 26). Arctic Sea Ice is not freezing In October for the first time since measurements began, now having an unknown effect on weather development towards Winter. *Severe Weather Europe*. https://www.severe-weather.eu/news/arctic-ocean-sea-ice-2020-jet-stream-effect-winter-fa/

Gertz, J. (2016). Reviewing METI: A critical analysis of the arguments. *ArXiv Preprint ArXiv:1605.05663*.

Ginsburg, I., Lingam, M., & Loeb, A. (2018). Galactic panspermia. *The Astrophysical Journal Letters*, *868*(1), L12.

Goff, J., Dominey-Howes, D., Chagué-Goff, C., & Courtney, C. (2010). Analysis of the Mahuika comet impact tsunami hypothesis. *Marine Geology*, *271*(3–4), 292–296.

Goldblatt, C., Robinson, T. D., Zahnle, K. J., & Crisp, D. (2013). Low simulated radiation limit for runaway greenhouse climates. *Nature Geoscience*, *6*(8), 661–667.

Goldblatt, C., & Watson, A. J. (2012). The runaway greenhouse: Implications for future climate change, geoengineering and planetary atmospheres. *Phil. Trans. R. Soc. A*, *370*(1974), 4197–4216. https://doi.org/10.1098/rsta.2012.0004

Gott III, J. R. (1993). Implications of the Copernican principle for our future prospects. *Nature*, *363*, 315–319.

Gowdy, J. (2020). Our hunter-gatherer future: Climate change, agriculture and uncivilization. *Futures*, *115*, 102488. https://doi.org/10.1016/j.futures.2019.102488

Grace, K. (2010). SIA doomsday: The filter is ahead | Meteuphoric. *Meteuphoric*. https://meteuphoric.com/2010/03/23/sia-doomsday-the-filter-is-ahead/

Greene, P. (2018). The Termination Risks of Simulation Science. *Erkenntnis*, 1–21.

Halstead, J. (2018). Stratospheric Aerosol Injection Research and Existential Risk. *Futures*. https://doi.org/10.1016/j.futures.2018.03.004

Handoh, I. C. (2013). Phosphorus and Chemical Pollution as Global Catastrophic Risks. *GCRI Blog*. http://gcrinstitute.org/handoh-lecture-summary-june-25-26/

Hansen, J., & et al. (1992).

Hanson, R. (1998). The great filter-are we almost past it. *Preprint Available at Http://Hanson. Gmu. Edu/Greatfilter. Html*.

Hanson, R., Martin, D., McCarter, C., & Paulson, J. (2021). A Simple Model of Grabby Aliens. *ArXiv Preprint ArXiv:2102.01522*.

Hartogh, P., Lis, D. C., Bockelée-Morvan, D., de Val-Borro, M., Biver, N., Küppers, M., Emprechtinger, M., Bergin, E. A., Crovisier, J., & Rengel, M. (2011). Ocean-like water in the Jupiter-family comet 103P/Hartley 2. *Nature*, *478*(7368), 218–220.

Henry, D. O. (2018). *From foraging to agriculture: The Levant at the end of the Ice Age*. University of Pennsylvania Press.

Herndon, J. M. (1993). Feasibility of a nuclear fission reactor at the center of the Earth as the energy source for the geomagnetic field. *Journal of Geomagnetism and Geoelectricity*, *45*(5), 423–437.

Hsu, D. C., Ford, E. B., Ragozzine, D., & Ashby, K. (2019). Occurrence Rates of Planets Orbiting FGK Stars: Combining Kepler DR25, Gaia DR2, and Bayesian Inference. *The Astronomical Journal*, *158*(3), 109. https://doi.org/10.3847/1538-3881/ab31ab

Hut, P., & Rees, M. J. (1983). How stable is our vacuum? *Nature*, *302*(5908), 508.

Karnaukhov, A. V. (2001). Role of the biosphere in the formation of the Earth’s Climate: The Greenhouse Catastrophe. *BIOPHYSICS-PERGAMON THEN MAIK NAUKA-C/C OF BIOFIZIKA*, *46*(6), 1078–1088.

Kelsey, H. M., Nelson, A. R., Hemphill-Haley, E., & Witter, R. C. (2005). Tsunami history of an Oregon coastal lake reveals a 4600 yr record of great earthquakes on the Cascadia subduction zone. *Geological Society of America Bulletin*, *117*(7–8), 1009–1032.

Kennett, J. P., Cannariato, K. G., Hendy, I. L., & Behl, R. J. (2003). *Methane Hydrates in Quaternary Climate Change: The Clathrate Gun Hypothesis*. American Geophysical Union.

Kent, A. (2004). A critical look at risk assessments for global catastrophes. *Risk Analysis*, *24*(1), 157–168.

Kipping, D. (2020). An objective Bayesian analysis of life’s early start and our late arrival. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1921655117

Kokaia, G., & Davies, M. B. (2019). *Stellar encounters with giant molecular clouds*. https://doi.org/10.1093/mnras/stz813

Kump, L. R., Pavlov, A., & Arthur, M. A. (2005). Massive release of hydrogen sulfide to the surface ocean and atmosphere during intervals of oceanic anoxia. *Geology*, *33*(5), 397–400.

Lenton, T. M. (2011). Early warning of climate tipping points. *Nature Climate Change*, *1*(4), 201.

Lingam, M., & Loeb, A. (2017). Risks for life on habitable planets from superflares of their host stars. *The Astrophysical Journal*, *848*(1), 41.

Lomborg, B. (2020). Welfare in the 21st century: Increasing development, reducing inequality, the impact of climate change, and the cost of climate policies. *Technological Forecasting and Social Change*, *156*, 119981. https://doi.org/10.1016/j.techfore.2020.119981

Lovelock, J., & Lovelock, J. E. (2000). *Gaia: A new look at life on earth*. Oxford Paperbacks.

Manheim, D. (2018). Questioning Estimates of Natural Pandemic Risk. *Health Security*, *16*(6), 381–390. https://doi.org/10.1089/hs.2018.0039

Marquis De Laplace, P. (1814). *A philosophical essay on probabilities*. Cosimo, Inc., 2007.

Napier, B., Asher, D., Bailey, M., & Steel, D. (2015). Centaurs as a hazard to civilization. *Astronomy & Geophysics*, *56*(6), 6–24.

Napier, W. M., & Clube, S. V. M. (1979). A theory of terrestrial catastrophism. *Nature*, *282*(5738), 455.

Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Hachette Books.

Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: Methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13*(2), 191–205.

Ozhovan, M. I., Gibb, F., Poluektov, P. P., & Emets, E. P. (2005). Probing of the interior layers of the Earth with self-sinking capsules. *Atomic Energy*, *99*(2), 556–562.

Popp, M., Schmidt, H., & Marotzke, J. (2016). Transition to a Moist Greenhouse with CO2 and solar forcing. *Nature Communications*, *7*. https://doi.org/10.1038/ncomms10627

Ramirez, R. M., Kopparapu, R. K., Lindner, V., & Kasting, J. F. (2014). Can increased atmospheric CO2 levels trigger a runaway greenhouse? *Astrobiology*, *14*(8), 714–731.

Ramirez, R. M., Lindner, V., & Kasting, J. F. (2013). How close is Earth to a runaway greenhouse? *ArXiv Preprint ArXiv:1306.5730*.

Rampino, M. R. (1998). The galactic theory of mass extinctions: An update. In *Dynamics of Comets and Asteroids and Their Role in Earth History* (pp. 49–58). Springer.

Rampino, M. R. (2008). Super-volcanism and other geophysical processes of catastrophic import. *Global Catastrophic Risks*, *1*, 203.

Rampino, M. R. (2015). Disc dark matter in the Galaxy and potential cycles of extraterrestrial impacts, mass extinctions and geological events. *Monthly Notices of the Royal Astronomical Society*, *448*(2), 1816–1820.

Rampino, M. R., & Caldeira, K. (2015). Periodic impact cratering and extinction events over the last 260 million years. *Monthly Notices of the Royal Astronomical Society*, *454*(4), 3480. https://doi.org/10.1093/mnras/stv2088

Reinhold, T., Shapiro, A. I., Solanki, S. K., Montet, B. T., Krivova, N. A., Cameron, R. H., & Amazo-Gómez, E. M. (2020). The Sun is less active than other solar-like stars. *Science*, *368*(6490), 518–521. https://doi.org/10.1126/science.aay3821

Sandberg, A. (2008). Bayes, Moravec and the LHC: Quantum Suicide, Subjective Probability and Conspiracies. *Andart*. http://www.aleph.se/andart/archives/2008/09/bayes\_moravec\_and\_the\_lhc\_quantum\_suicide\_subjective\_probability\_and\_conspiracies.html

Sandberg, A., Drexler, E., & Ord, T. (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

Sandberg, A., Snyder-Beattie, A., & Armstrong, S. (2018). Nuclear war near misses and anthropic shadows. *Manuscript in Preparation*.

Seeley, J., & Wordsworth, R. (2021). Episodic deluges in simulated hothouse climates. *Nature*, *599*(7883), 74–79. https://doi.org/10.1038/s41586-021-03919-z

Segschneider, J., Beitsch, A., Timmreck, C., Brovkin, V., Ilyina, T., Jungclaus, J. H., Lorenz, S. J., Six, K. D., & Zanchettin, D. (2013). Impact of an extremely large magnitude volcanic eruption on the global climate and carbon cycle estimated from ensemble Earth System Model simulations. *Biogeosciences*, *10*, 669–687.

Shakhova, N., Semiletov, I., Salyuk, A., Yusupov, V., Kosmach, D., & Gustafsson, Ö. (2010). Extensive methane venting to the atmosphere from sediments of the East Siberian Arctic Shelf. *Science*, *327*(5970), 1246–1250.

Shcherbakov, А. С. (1999). Антропный принцип в космологии и геологии. *Вестник Московского Университета. Серия*, *7*, 58–70.

Snyder-Beattie, A. E., Ord, T., & Bonsall, M. B. (2019). An upper bound for the background rate of human extinction. *Scientific Reports*, *9*(1), 11054. https://doi.org/10.1038/s41598-019-47540-7

Snyder-Beattie, A. E., Sandberg, A., Drexler, K. E., & Bonsall, M. B. (2020). The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare. *Astrobiology*. https://doi.org/10.1089/ast.2019.2149

Song, H., Kemp, D. B., Tian, L., Chu, D., Song, H., & Dai, X. (2021). Thresholds of temperature change for mass extinctions. *Nature Communications*, *12*(1).

Strom, R. G., & Schaber, G. G. (1992). Pulsed Resurfacing Events on Venus, Earth and Mars. *Bulletin of the American Astronomical Society*, *24*, 946.

Strom, R. G., Schaber, G. G., & Dawson, D. D. (1994). The global resurfacing of Venus. *Journal of Geophysical Research: Planets*, *99*(E5), 10899–10926.

Tegmark, M., & Bostrom, N. (2005). How unlikely is a doomsday catastrophe? *ArXiv:Astro-Ph/0512204*. http://arxiv.org/abs/astro-ph/0512204

Thrasymachus, K. (2012). *UFAI cannot be the Great Filter*. https://www.lesswrong.com/posts/BQ4KLnmB7tcAZLNfm/ufai-cannot-be-the-great-filter

Totani, T. (2020). Emergence of life in an inflationary universe. *Scientific Reports*, *10*(1), 1671. https://doi.org/10.1038/s41598-020-58060-0

Trigo-Rodríguez, J. M., & Williams, I. P. (2017). Dynamic sources of contemporary hazard from meteoroids and small asteroids. In *Assessment and Mitigation of Asteroid Impact Hazards* (pp. 11–32). Springer.

Turchin, A. (2018a). *A Meta-Doomsday Argument: Uncertainty About the Validity of the Probabilistic Prediction of the End of the World*. https://philpapers.org/rec/TURAMA-4

Turchin, A. (2018b). *Forever and Again: Necessary Conditions for the “Quantum Immortality” and its Practical Implications*.

Turchin, A. (2018c). The Risks Connected with Possibility of Finding Alien AI Code During SETI. *Journal of British Interplanetary Society*, *70*.

Turchin, A. (2020). *Presumptious philosopher proves panspermia*.

Turchin, A., & Denkenberger, D. (2018). Global Catastrophic and Existential Risks Communication Scale. *Futures*, *102*, 27–38.

Turchin, A., & Denkenberger, D. (2019). Classfication of ETI riks. *Inder Review in JBIS*. https://philpapers.org/rec/TURGCR

Turchin, A., Yampolskiy, R., Denkenberger, D., & Batin, M. (2019). *Simulation Typology and Termination Risks*.

Turner, M. G., Romme, W. H., & Tinker, D. B. (2003). Surprises and lessons from the 1988 Yellowstone fires. *Frontiers in Ecology and the Environment*, *1*(7), 351–358.

Tyrrell, T. (2020). Chance played a role in determining whether Earth stayed habitable. *Communications Earth & Environment*, *1*(1), 1–10. https://doi.org/10.1038/s43247-020-00057-8

USGS. (2020). Volcanoes Can Affect Climate. *USGS*. https://www.usgs.gov/natural-hazards/volcano-hazards/volcanoes-can-affect-climate

Waltham, D. (2011). Testing anthropic selection: A climate change example. *Astrobiology*, *11*(2), 105–114. https://doi.org/10.1089/ast.2010.0475

Waltham, D. (2014). On the absence of solar evolution-driven warming through the Phanerozoic. *Terra Nova*, *26*(4), 282–286.

Waltham, D. (2019). Is Earth special? *Earth-Science Reviews*.

Ward, P. D. (2007). *Under a green sky*. Smithsonian Books/Collins.

Ward, P. D., & Brownlee, D. (2003). *Rare Earth: Why Complex Life is Uncommon in the Universe*. Copernicus.

Ward, P. D., & Brownlee, D. (2004). *The life and death of planet Earth: How the new science of astrobiology charts the ultimate fate of our world*. Macmillan.

Watson, A. J. (2004). Gaia and observer self-selection. *Scientists Debate Gaia: The next Century*, 201–210.

Wordsworth, R. (2021). How likely are Snowball episodes near the inner edge of the habitable zone? *ArXiv:2104.06216 [Astro-Ph]*. http://arxiv.org/abs/2104.06216

Yudkowsky, E., & Soares, N. (2017). Functional Decision Theory: A New Theory of Instrumental Rationality. *ArXiv:1710.05060 [Cs]*. http://arxiv.org/abs/1710.05060

# Appendix 1. Different types of anthropic shadow

[appendixes are not proofread yet- AT]

(Ćirković et al., 2010) did not discuss how past catastrophes may affect the rate of future catastrophes; it was assumed in that work that each catastrophe is a truly random event, similar to radioactive decay, where past and future events are mutually independent. This is the simplest case, but not all natural catastrophes are mutually independent events. Rate estimation for non-mutually independent events is also affected by the anthropic shadow and could give even worse expectations of future rates after the anthropic shadow is taken into account.

## Random events and regression to the mean

The first category of natural catastrophes comprises completely random and mutually independent events, for example, a supernova explosion not far from Earth (Ćirković & Vukotić, 2016). The anthropic shadow here creates an underestimation of the natural rate of such events, and after the current moment, there will be a regression to the mean of this rate, seen here as a return to a *higher* probability of this type of catastrophes.

## Periodic catastrophes

The *Rare Earth Hypothesis* (REH), suggested by P. Ward (Ward & Brownlee, 2003), asserts that the origin of life and the evolution of biological complexity – including multicellular organisms and human intelligence – is a result of an incredibly unlikely combination of astrophysical and geological events and circumstances and thus the Earth is probably the only habitable planet in the observable Universe. If the REH is true, there were many *Early Filters* – or one very strong – which the Earth had to pass through, and for each there was only a small probability that the Earth was actually able to pass through it.

The *Great Filter* is a hypothetical event which prevents observations of other alien civilizations and thus is a solution of the Fermi paradox (Hanson, 1998). There are two types of such filters: Early Filters, which affect planets before current-level-civilization arises, and Late Filters, which could happen for us in the future. Early filters themselves could be either *positive*, like events which should happen for civilization to evolve: abiogenesis, multicellularity, sexual reproduction; and *negative*: the events *which should not happen*, that is large asteroid impacts, large solar flares, close supernovas. Negative early filters are mostly catastrophes.

In the real world, there are very few strictly periodic catastrophes. However, it has been suggested that the Solar system has oscillated relative to the galactic plane while it is moving around the galactic center with a period around 30 million years—the *Shiva hypothesis* (Napier & Clube, 1979).

Such passing through the galactic plane, which has a higher concentration of matter (including probably dark matter), could disturb objects in the Oort cloud, which would then start falling towards the Sun and increase the rate of dangerous comets in the inner Solar system, including invisible *dark comets*. However, it would take up to a few million years for the disturbed comets to reach the inner Solar system, as they fall very slow initially. At first, they become Centaur-type bodies beyond Jupiter’s orbit (Napier, Asher, Bailey, & Steel, 2015), then go closer to the Sun, break up after flying by the Sun and only after that they become dangerous swarms of dark kilometer-sized bodies with short periodic orbits. For instance, the Encke comet remains (Trigo-Rodríguez & Williams, 2017) are a probable progenitor of the Tungska event impactor. There is a school of thought suggesting that Earth is currently in a period of more frequent asteroid bombardment than the median, including some recent events as the Clovis comet in 12,900 BC (Firestone et al., 2007) and the Mahuika crater progenitor in the 15th century (Goff et al., 2010). Moreover, it was suggested that passing through the galactic plane could heat interior of the Earth via dark matter annihilation (Rampino, 2015) which would explain the simultaneous volcanic activity.

The last mass extinction event was 65 million years ago, or around two galactic plane crossings ago (Rampino, 1998) and there was a smaller extinction peak (Eocene–Oligocene extinction event, which coincides with two large impacts, including Popigai crater) around 33.9 millions years ago. Rampino wrote: “Statistical analyses suggest that mass extinction events exhibit a periodic component of about 30 Myr, and periodicities of 30 ± 0.5 Myr and 35 ± 2 Myr have been extracted from sets of well-dated large impact craters. These results suggest periodic or quasi-periodic showers of impactors, probably Oort Cloud comets, with an approximately 30 or 36 Myr cycle”.

The 30 million years cycle is long overdue and the 35 ± 2 cycle could already have started, which implies that the next large impact is probable in 1-2 million years from now, but not in 30 Myr. (There are theories that we already living in the period of more intense asteroid bombardment than average, and there were several impacts in recent historical time, including Clovis comet 12.7 ky from now; the analysis of validity of these claims is out of scope of this paper.)

From rather general non-anthropic considerations, any complex life is more likely to appear *closer to the end of a period of stability*, just because at the beginning of such period everything is in ruins. The main feature of the quasi-periodic catastrophic events is that the future probability of such an event is not independent from its past probability, and the longer such event did not happen, the more probable it becomes. This is not true for comets, but more probable for supervolcanos, which accumulate energy for the next eruption. Our survival is the evidence that any periodic events which preclude our existence did not happen for a long time.

We could define *anthropic shadow as a survivorship bias applied to our own existence*. We can’t observe evidence of past catastrophes which, for sure, would cause our extinction. But we could observe catastrophes which were smaller and only had some probability of destroying life on Earth. For example, there is no evidence of impacts of 100 km size body on Earth for last billion years. However, smaller 10-20 km bodies did impact us. But it is incorrect to use their impact rate to predict future impact rate as it is distorted by survivorship bias. We need to take the magnitude of that bias into account.

Ćirković et al. suggested measuring the anthropic shadow using the *overconfidence parameter η*, which is the relationship of the *a* *priori* and *a posteriori* probability of a catastrophe, where the *a posteriori* probability is the probability updated with an account of the anthropic shadow (Ćirković et al., 2010). To put this into anthropic terms, we need to state the notion of the *anthropic pressure,* which is the share of typical earth-like planets that survives a given filter.

If there is a periodic catastrophe, and we know the period, we should expect the catastrophe in the remaining part of the period, not the whole period. For example, if there is 35 ± 2 Myr period in asteroid bombardment, but only 34 Myr passed from the last such period, we should expect the next bombardment in 0-3 Myr from now, which implies *order of magnitude* greater catastrophe’s probability than for truly random event.

If there is no anthropic pressure, a random observer will be somewhere in the middle of the period between two periodic catastrophes. However, if there is anthropic pressure with overconfidence parameter *η,* then the most likely location of the now moment is 1/η of the time until the next periodic catastrophe. For example, if the period is 30 million years and η =10, the next periodic catastrophe is likely in 3 million years.

## Semi-periodic catastrophes, which accumulate and then release energy

Another type of periodic catastrophes is those which accumulate energy, like a tensioned spring. The more a spring is elongated, the more energy it accumulates, and the more probable its catastrophic release is.

Forest fires (Turner et al., 2003) and megathrust earthquakes are examples of this type of catastrophes, though they are not global catastrophes. Nonetheless, they accumulate and then release energy; in the case of megaquakes like the Cascadia earthquake, they happen every few hundred years, and the last happened in 1700 (Kelsey et al., 2005).

One possible scenario of a global catastrophe which requires the accumulation of energy is a basalt flow eruption, e.g. the Siberian traps eruptions, the probable cause of the Permian-Triassic extinction event, the most severe extinction event in Earth’s history, some 252 million years ago (Carpenter & Bishop, 2009). Another is a possible catastrophic release of methane hydrates from the ocean floor via a positive temperature feedback loop (this is a more speculative event). Some claim that there are currently 10 times more methane hydrates on the ocean floor than during the Paleocene–Eocene Thermal Maximum (PETM) temperature excursion 55 million years ago, when temperature jumped around 8 °C for around 200 000 years (Bowen et al., 2015).

However, the most severe possible catastrophes may never have happened in Earth’s history. For example, Venus has experienced complete surface recycling (melting) after enormous lava event around 1 billion years ago, based on its low cratering (Strom et al., 1994): there are no plate tectonics on Venus, and no ways to gradually release internal energy; it is assumed that during surface replacement event the whole old surface was covered by lava. Earth may also have had pulses of more active plate tectonics, resulting in effective resurfacing of most of the crust (Strom & Schaber, 1992).

# Appendix 2. Normally distributed anthropic shadow

We could start with the following metaphor: Imagine that we live on a surface of a balloon which is slowly inflating. Its explosion is long overdue, but we don’t know this, as we have never observed other balloons. However, the balloon most likely can be inflated a couple of more centimeters. But if we start to poke into its surface, it will explode immediately.

## Toy example with an old man

If a catastrophe is inevitable, but its timing is *normally* distributed around some mean value, and we have survived n-sigmas after that event, our probability to survive until n+1 sigma is not high, but we can’t know it because of the survivorship bias.

Let’s create a toy example with human life expectancy, where it is distributed normally, but not according to the double exponential Gompertz law, as in reality. After we illustrate the main idea, we could apply it to other types of distributions as we do Appendix 2.

In our toy example, a 90-years old person lives alone on an island and does not know the laws of aging, but can observe that he is aging. He has a mean life expectancy of 80 years and the death probability is normally distributed around the age of 80 with a standard deviation of 10 years. The questions we are interested in is: what is his *remaining life expectancy* in 90 years old and how his life expectancy is different from his naive expectations? We will show below in the box that:

*1. Even a minor anthropic shadow means a dramatic cut of remaining life expectancy*. If he does not account for anthropic shadow, he should expect to live 90 more years, but if he would take into account anthropic shadow, he will have only 3.58 years life expectancy.

*2. However, the sensitivity of the remaining life expectancy to the power of anthropic shadow is small because it is calculated via conditional probability and logarithm.* For 90yo it is 3.58 years life expectancy. If he will be 100 years old person, he will have 2.43 years life expectancy. And for 110 years it will be 1.837 years. His remaining life expectancy declines slowly (around two times) while his chances to survive to each subsequent decade decline dramatically: 15.9% for 90, then 2.3% for 100 years and then 0.135% for 110 years.

*3. Any anthropic shadow means significant growth of fragility.* The old man’s fragility has grown, which could be estimated as a height, jumping from which could be fatal. For real humans, it declines from single meters in young age, to tens centimeters in very old age, or around one order of magnitude.

*4. Powerful anthropic shadow is unlikely as most old observers will find themselves in 1 and 2 sigma periods*. The old man is unlikely to be very old, as most old men are below 100 years old (more in section 5). Therefore, we could ignore higher sigmas and look only at the reaming life expectancy for 90 or 100 years old man, and it is 3.58 or 2.43, or around 3 years.

5. *In most real cases, future life expectancy is between 35.8 and 12.15 per cent of the known past life if anthropic shadow is present*. In the most cases, we don’t know the power of anthropic shadow, but we do know the past time of the existence and we can estimate the ratio between future life expectancy and past life expectancy. We need to do it, because future life expectancy above was measured in sigmas, but we do not know how many sigmas are in the past time of existence, so we need to convert sigmas in years. For old man at 90 this ratio is 3.58:10 = 0.358, and for 100yo it is 2.43: 20= 0.1215. (For the old man, past life is calculated starting from 80 in our toy example, as it is the time of the mean.) We ignore higher sigmas as they are unlikely. For 3 sigma, the ratio is 6.12 % which is also close to 1 order of magnitude, but the chances of having such strong anthropic shadow is only 2.1 per cent.

6. Over simplification of all said above is the following rule of thumb: *Anthropic shadow lowers future life expectancy for no more than 1 order of magnitude in most plausible cases*. As we do risks analysis here, we have to take the upper level of expected risk, so we can say that anthropic shadow lowers life expectancy for one order of magnitude.

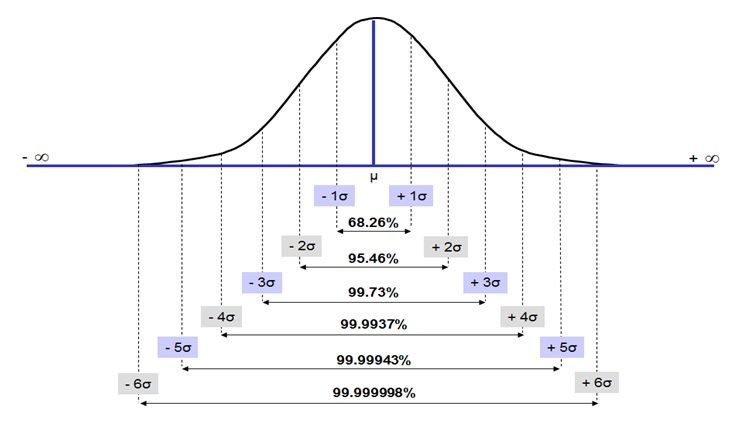
7. In the cases where the fragility is proportional to the future life expectancy, like for an inflating balloon, the presence of *anthropic shadow means an order of magnitude increase of fragility.*

In the box, we delve into the math of our toy example and we will dig it even deeper in Appendix 1, but the reader could jump to next section, where the central argument is presented.

## Table 2. Life expectancy of survival if normally distributed catastrophe is long overdue.

Relation between normal distribution’s sigmas and the life expectancy could be illustrated by the following table where values are numerically calculated. General solution will be in the next sub-section.

The interesting feature of this table is that while the survival probability declines very quickly, the *expected survival half-life* expressed in time units is declining *logarithmically* and relatively slow (in the second from right column).



The chances to survive and until the “mean age plus one sigma” is 100 – (50+34.1) ≈ 15.9 %.

The chances to survive until “mean time + 2 sigma” are 100 – (50+34.1+13.6) ≈ 2.3 %,

The chances until 3 sigma it is 100 – (50+34.1+13.6 +2.1) ≈ 0.135%.

And until 4th sigma it is: 100 – (50+34.1+13.6 +2.1+0.0063) ≈ 0.00315%.

As we are interested in the anthropic shadow, we should look at the *conditional probability*: what are our chances to survive until n+1 sigma, given that we have survived until the moment of *n* sigma. As the chance to survive until 1st sigma is 15.9% and to 2nd sigma is 2.3%, so the conditional probability is 2.3:15.9 ≈ 0.144=14.4%. The conditional probability to survive from 2nd sigma to 3d sigma is 0.135:2.3 ≈ 0.058=5.8%. The chance to survive to 4th sigma is 0.00315:0.135 = 0.023 = 2.3%.

The first interval, from 1th sigma to 2nd sigma had 0.144 survival chance. The number of “doublings of risk” or “halvings of the survival probability” in that period is log0.5 (0.144) ≈ 2.79 (given our linearity assumption).

Therefore, life expectancy after 1st sigma is 1: 2.79 = 0.358 sigma, or 3.58 years for our old man. The same way, the life expectancy after 2nd sigma is 10:log0.5 (0.058) ≈ 10:4.10 ≈ 2.43 years.

For 3 sigma, it is 10:log0.5 (0.023) ≈ 1.837 years

### Box. The analysis of the toy example

In Figure 1 we see normally distributed *timing* of some inevitable catastrophe: the probability that the catastrophe has happened before the moment *t* is a sum of all probabilities between -∞ and *t*. In that case, the chances to survive until the mean age are 50 %.

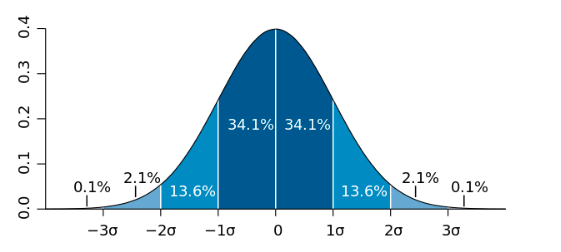
The chances to survive and until the “mean age plus one sigma” is 100 – (50+34.1) ≈ 15.9 %.

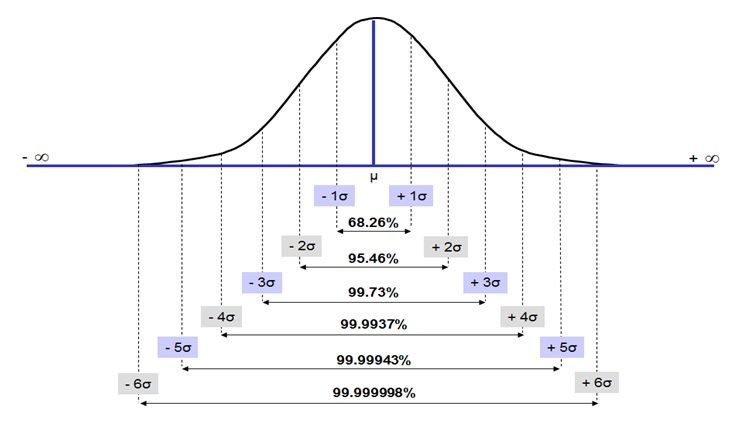
The chances to survive until “mean time + 2 sigma” are 100 – (50+34.1+13.6) ≈ 2.3 %,

The chances until 3 sigma it is 100 – (50+34.1+13.6 +2.1) ≈ 0.135%.

And until 4th sigma it is: 100 – (50+34.1+13.6 +2.1+0.0063) ≈ 0.00315%.

*Figure 1. Normal distribution of the timing of an inevitable catastrophe from the mean.*





As we are interested in the anthropic shadow, we should look at the *conditional probability*: what are our chances to survive until n+1 sigma, given that we have survived until the moment of *n* sigma. As the chance to survive until 1st sigma is 15.9% and to 2nd sigma is 2.3%, so the conditional probability is 2.3:15.9 ≈ 0.144=14.4%. The conditional probability to survive from 2nd sigma to 3d sigma is 0.135:2.3 ≈ 0.058=5.8%. The chance to survive to 4th sigma is 0.00315:0.135 = 0.023 = 2.3%.

Therefore, chances to survive to the next sigma are declining: after first sigma, it is 14%, after second, it is 5.8% and after third, it is 2.3%. (See more results in Table 1 in Appendix 1.)

From above follows that for the first two sigmas the probability of reaching the next sigma is around 10%. We will now show that it is unlikely that anthropic shadow is strong, based on anthropic arguments, so we should ignore higher sigmas. In our toy example, if a person finds that he is physically old, he concludes that he is older than the mean age of death, but for how many sigmas? He could use *anthropic logic*: for 100 random people, 50 already dead, 34 are in the first sigma, 13 in second and 2 in third and less than one in highest sigmas.

Thus, he has like 34:50 = 68% to be in the first sigma period of the distribution and 26% in the second, and only 6 in higher sigmas. So, it is unlikely to be an observer who observes strong anthropic shadow. As we explain in section 4, this objection can’t kill anthropic shadow completely as there is a *minimal level* of its power, and in our case the minimal level is that the old man knows that he is old and he is older than the mean age, so he is at least in the first sigma. From this follows that he could ignore any hypothesis that he is in sigmas higher than 2. And we showed above that for the 1 and 2 sigmas, the chances to survive to the next sigma is around 10 per cent.

Therefore, the old man could reasonably bet that he can survive the next 10 years with around 10 per cent chances.

Similar to the case of human life expectancy, it is better to speak about *remaining* *life expectancy* than of the “probability to survive until the next sigma”. We define “life expectancy” as a time from now until the moment when the probability of survival will fall to 0.5. (As the function declines rather monotonically between larger sigmas, we could approximate it here by a linear function between nearest sigmas, which will help us to estimate life-expectancy; more general solution will be shown in Appendix 1.)

The first interval, from 1th sigma to 2nd sigma had 0.144 survival chance. The number of “doublings of risk” or “halvings of the survival probability” in that period is log0.5 (0.144) ≈ 2.79 (given our linearity assumption).

Therefore, life expectancy after 1st sigma is 1: 2.79 = 0.358 sigma, or 3.58 years for our old man. The same way, the life expectancy after 2nd sigma is 10:log0.5 (0.058) ≈ 10:4.10 ≈ 2.43 years.

For 3 sigma, it is 10:log0.5 (0.023) ≈ 1.837 years.

For Earth history, we know all time of existence, but do not know how many sigmas are in it. Therefore, we need *Ratio of the future expected survival time – to the past time* of existence without catastrophes, if we want to calculate future life expectancy, and it is: 35.8% for 1 sigma, 12.5% for 2 sigmas, and 6.12 % for 3 sigmas (see more in Table 1 in Appendix).

Now, if we apply the above calculations to our toy example, it means that a 90-years-old person has 3.58 years life expectancy and a 100-years-old person has 2.84 years life expectancy. (In real life, Gompertz law is steeper and 100 years old person has around 1 year life expectancy.) We could be surprised to see that the remaining life expectancy didn’t fall that much despite the fact that total chances to survive to 100 years are more than 10 times less than the chances to survive to 90 years.

For both 1 and 2 sigmas future life expectancy is around 30 per cent of the sigma. Because life expectancy is defined via logarithm, it is changes *slowly* than probability to survive to the next sigma. This slow change could be interpenetrated as *low sensitivity of the anthropic shadow to initial parameters*. Especially if we cut higher sigmas based anthropic considerations.

Now, imagine that the person from our thought experiment lived all his life on island and does not have *a priory* knowledge about typical human life expectancy. He only knows that he is 90 years old. He may use the Laplace’s sunrise problem’s solution and estimate that his future life expectancy is around 90 more years.

However, he learns from the observation of some animals that they die from aging, but he still does not know exactly what is the mean human life expectancy, except that it is in the past for him. It could be 70 or 80 years. In that case, he could apply our calculations from above to get that his future life expectancy is around 2 – 3 years. He also could ignore decimal digits after 2 and 3 with all uncertainties about his exact timing of human mean life expectancy and the shift of the size of the sigmas.

Moreover, he should conclude that his frailty has increase. In the past, he could jump from 4 meters rock, but now a fall on a plain spot could be fatal. If aging complexly explained by frailty, his decline of life expectancy will be proportional to the decline of the “deadly height” from which he can safely jump. If life expectancy declined for an order of magnitude, then, at first approximation, the frailty has increased around 10 times. So, it will be not 4 meters, but only 40 cm.

*2. However, the sensitivity of the remaining life expectancy to the power of anthropic shadow is small.* If he will be 100 years old person, he will have 2.43 years life expectancy. And for 110 years it will be 1.837 years. His remaining life expectancy declines slowly (two times) while his chances to survive to each age decline dramatically: 15.9% for 90, then 2.3% for 100 years and then 0.135% for 110 years.

For Earth history, we know all time of existence, but do not know how many sigmas are in it. Therfore, we need *Ratio of the future expected survival time – to the past time* of existence without catastrophes, if we want to calculate future life expectancy, and it is: 35.8% for 1 sigma, 12.5% for 2 sigmas, and 6.12 % for 3 sigmas (see more in Table 2 in Appendix).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *Sigma deviation* from the mean, n | *Chances to survive to sigma = n (****after*** *the mean value, so we divide on half)* | *Conditional probability to survive until sigma = n+1*, given that we already survived until sigma = n: | Number of halvings of the probability in this interval to get the value in previous column | *Half-life survival time* from now | *Ratio of the future expected survival time – to the past time* of existence without catastrophes |
| n = 1 | 1 –(0.341+0.5)  = 0.159  1 in 6 (which is 0.5 of [typical](https://en.wikipedia.org/wiki/Normal_distribution#Standard_deviation_and_coverage) 1 in 3 because we take only the right part of the distribution) | 2.3: 15.9 = 0.144. | log0.5 (0.144) = 2.79 halvings | 1:2.79 = 0.358 sigma | 0.358:1= 35.8% |
| n = 2 | 100– (50+34.1+13.6) ≈ 2.3 %  1 in 43 | 0.135:2.3 ≈ 0.058 | log0.5 (0.058) ≈ 4.1 doubling | 0.243  sigma | 0.243:2=  12.5% |
| n = 3 | 100 – (50+34.1+13.6 +2.1) ≈ 0.135%.  1 in 740 | 740:31574 =  1 in 42.66  = 0.0234 | log0.5 (0.023) ≈ 5.44 doublings | 0.1837 sigma | 0.1837:3 ≈  6.12 % |
| n = 4 | 1 in 31 574 | 31 574: 3 488 556 = 1 in 110.48… | ≈ 6.8 doublings | 0.14 sigma | ≈ 3.5 % |
| n = 5 | 1 in 3 488 556 | 1 in 290… | ≈ 8.2 doublings | 0.121 sigma | ≈2.42 % |
| n = 6 | 1 in  1 013 594 692 | 1 in 770.88 | ≈ 9.6 doublings | 0.104 sigma | ≈ 1.73 % |
| n=7 | |  |  | | --- | --- | | 0.5 in | 390682215445 | |  |  |  |  |
| n= x | Erf(x/sqrt2)) |  |  |  |  |
| n=14 | 1 in 1045 |  | 150 doublings | 1:150 sigma | (1:150):14=  0.0004 = 0.04 per cent |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

*Table 1. Life expectancy of survival if normally distributed catastrophe is long overdue.*

## Mathematical model

Here we will calculate how future survival time depends on the power of the anthropic shadow in the case of a semiperiodic event which has normally distributed deviation from the period.

In the case of the normal distribution, the probability of extinction to time *t* (calculated after the period time Tp) is given by a *cumulative distribution function*:

=

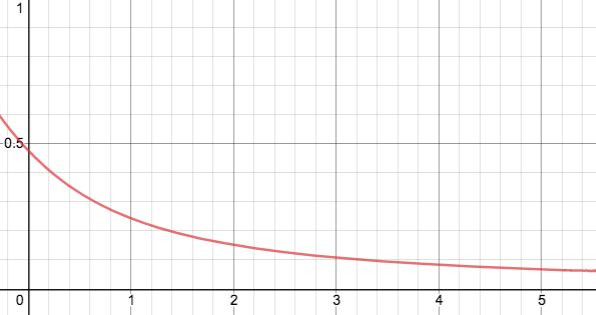
If humanity already survived until the moment T, its chances of surviving until the moment T+∆t are:

S=

We could present the anthropic pressure as a half-life expectancy: how long from now will there be a catastrophe with 50 percent chance? To estimate median life expectancy, we need to calculate time until the total probability of survival falls 2 times.

P(tnow+thalf-life) = 0.5 P(tnow)

This equation has to be solved for thalf-life, but it cannot be solved algebraically, so we solve it numerically for different P(tnow) by creating a chart of the equation *1-erf(x) =2(1-erf(x+y))*



y, expected half-life in the moment x

x, time in sigma deviations from median life expectancy

*Chart 1. In this chart, x is normalized time of the current deviation from the median time of a periodic catastrophe (correlated with anthropic bias), and y is the expected half-life at the moment x, given the anthropic shadow.*

For x=1, it gives y≈0.25, for x=2, it gives y≈0.15 and for x = 3, y≈0.1. For example, if the current deviation from the expected time of semi-periodic asteroid impact is 2 million years (with 1-million-year standard deviation), the remaining time until the next impact with 50 percent probability is 150 000 years, according to Chart 1, and if now-moment is at 4 million years deviation, the life expectancy is 90 000 years.

The interesting feature of this solution is that it is very *flat* after 1, and values for y are around 0.1 for x in (2;5) despite the fact that surviving until 4 sigma deviation is much less probable than 2 sigma deviation. This means that our future anthropic life expectancy does not depend (at first approximation) on the power of the anthropic shadow, and is around 0.1 of standard deviation.

The interval of approximately (2, 5) is where the anthropic shadow works. Below 2 is just normal deviation of random variable, and there is small anthropic shadow. But the anthropic shadow cannot be very strong because of the *self-indication-counterargument* which was discussed in section 4.1, which implies that one is more likely to find oneself in the world with *weaker* anthropic shadow, as such worlds have more observers. Thus, the anthropic shadow works only for a rather short interval of possible deviations, the actual boundaries of which are context dependent. These boundaries are not very important, as predicted life expectancy varies within them from 0.15 to 0.09, but given the large uncertainty of the period and the anthropic shadow, only the first digit is relevant, so it could be estimated at around 0.1 sigma.

To convert this 0.1 sigma life expectancy into time units, we need some estimates of the variability of a semi-periodic catastrophe. If the catastrophe is strictly periodic, sigma is equal 0, and the next event is soon. If it is a one-time event, it could be seen as normally distributed from the beginning, without a periodic component. In that case, we could use the whole previous period of stability as *x* from Chart 1. In that case, we know only the past time, but not the anthropic shift or sigma. But as *y is not sensitive to initial parameters, we could calculate that* *y/x* is part of an interval (0.085; 0.03; 0.02; 0.014) for x in (2, 3, 4, 5) respectively.

In other words, typically one can guess neither a catastrophe probability, nor the magnitude of the anthropic shift, but we have only duration of the past stable period and some evidence that there is an anthropic bias. But such duration is enough to make a very rough estimate of future survival time, which we above estimated as from 1.4 per cent to 8.5 per cent of the past survival time, and closer to the higher level of 8.5 per cent, because of SIA counterargument. Oversimplifying, we could state that anthropic bias means that future survival time is order of magnitude less than past survival time.

## Life expectancy based on anthropic bias is approximately the same for different catastrophe distributions

In real life, one unlikely knows for sure if some type of global catastrophe is periodic, normally distributed and-or long overdue. Or, if one knows everything about this type of catastrophe, there will be no place for any anthropic bias. Anthropic bias appears in the situation where one knows little about the real nature of the catastrophe, except its past frequency.

The analysis above that if there is some evidence for anthropic bias, then future life expectancy will be around 0.1 of past stable time. This claim is much stronger than traditional Doomsday argument, but it works in the situations where we expect strong observer selection effects resulting in underestimation of the rate of past catastrophes. For some catastrophe types, it could be larger or smaller, but it is compensated by our uncertainty in which type of catastrophe we observer (or do not observe). The estimation of “one order of magnitude lower remaining time” is itself very approximate, and it just means that it is not two orders of magnitude reduction. However, as we showed in the article which explores the problem of communicating and measuring probability of global risks (Turchin & Denkenberger, 2018), in most cases, society does not need more detailed information than just an order of magnitude of the probability of the catastrophe, because it will not significantly change the prevention plans.

One may conclude that based on this risk correction, one could safely continue to ignore most of the natural global risks, as one order magnitude of increase of their probability is not enough to make them significant compared to known anthropogenic risks. For example, if civilization-ending asteroid impact may happen not in ~30 million years from now, but in ~3 million years, it still means chances of 0.003 per cent in next century, which is significantly smaller than most conservative estimates of nuclear war and other anthropogenic risks.

But the problem is not closed, as the survivorship bias makes the world more *fragile* to the anthropogenic risks, as we will discuss in the next section.

The higher is the anthropic pressure (the magnitude of underestimation of past risks because of anthropic shadow), the lower is the threshold for the tipping point and the smaller are the human actions that could cause a change to another stable state. However, the anthropic pressure cannot be too high, as alternative worlds with lower anthropic pressure would be more numerous as we will discuss in Section 4.1.

There are other possible positive feedbacks (besides methane increase) including different chemicals in upper atmosphere and reversal of *global diming* because of coal usage decline and global lockdown, which could trigger runaway warming and which potential could be underestimated. In October 2020 Arctic was +6C warmer than average and ice was growing extremely slow (Flis, 2020).

# Appendix 3. Other ways to estimate the power of anthropic shadow

## 5.4. Anthropic shadow magnitude could be estimated by frequency of near-misses

A “near-miss” is a situation when a catastrophe nearly happened, but did not. An example situation is when a car has to use emergency braking, that is, all but one safety system fails. For known technical systems, the frequency of near-misses is around 50-100 near-misses for one actual accident (Bridges, 2012). (There is some uncertainty in the “near-miss” definition, as some dangerous situations could pass unnoticed. Also, it was found that one near-miss is happened for every 100 rules violations.)

If one uses the estimate of 100 near-misses per catastrophe, one could roughly estimate the probability of catastrophes which nearly happened. For example, around 15 near-misses of nuclear war are publicly accepted, including the Caribbean crisis and Petrov’s accident [Wikipedia]. (Unfortunately, such incidents remain secret for a long time and there are rumors of other incidents, so the actual number is not accessible and could very well be several times higher.) 15 near-misses is well below 100, so it appears that there is little anthropic pressure for nuclear war observations. The rate of publicly acknowledged nuclear near-misses is one per 5 years, which implies median time of nuclear war is 500 years. But if one assumes that only half of nuclear near misses is reported, and that low bound of the rate is 1 to 50, nuclear war would happen once in 125 years, and anthropic shadow for it is 125:75= 1.66.

Applying near-misses concept to past natural catastrophes is more difficult, as there are little data on events which did not reach the required level of severity. The five major extinctions were probably near-misses for killing multicellular life. For more recent times, a near-miss could have been the Toba eruption and perhaps the Clovis comet, if real. For climate change, the small number of near misses is not so obvious, as the Earth several times was close to snowball state, but only once close to runaway warming (PETM). However, here, as in the famous example of survival ship bias with holes in returning warplanes, we should look at the events which are not observed: and the lack of extreme warming periods could be indication that such warming would be fatal.

## 5.5. Testing the “anthropic shadow” on the known catastrophes

From the calculations above follows an important conclusion: *one does not need to know the magnitude of survivorship bias to roughly estimate the future half-life expectancy*: future half-life will be approximately the same for all reasonable values of the survivorship bias, that is, for 1-2 sigmas, that is, 1 in 43 (see Table 2 in Appendix).

In other words, if one knows that survivorship bias is in play, and the age of the observed object is T, the approximate future *life expectancy estimation based on the anthropic shadow* *is around 0.1T*. This type of prediction is similar to the Gott’s Doomsday Argument (DA), but for DA the future *life expectancy is T, the same as past one.*

This observation may work in the opposite direction: if we know the past and future life expectancy, and the future life expectancy is around 0.1 of the past, it is evidence that there was an anthropic pressure, but it *is not evidence that past anthropic pressure was only 10 per cent*! Actually, it is likely no more than 2.3 per cent, because observing such significant anthropic shadow is unlikely, except the situation of very high base rate of catastrophes over all possible planets, as will be discussed in section 5. It also means that it is very unlikely to observe the 0.01T as a future survival time and such observation is likely to be wrong. However, the article “The Timing of Evolutionary Transitions Suggests Intelligent Life Is Rare” by Sandberg et al (Snyder-Beattie et al., 2020) shows that the base rate for positive fine-tuning could be extremely high, many orders of magnitude, and as we assume that positive and negative fine tuning have approximately the same power, it would mean very high base rate of catastrophes.

The idea that there was a strong anthropics shadow on Earth can be checked. Life on Earth has existed for around 3.8 billion years, and the future life expectancy of biological life on Earth before the growth of Sun’ luminosity will evaporate the oceans is estimated to happen 1 billion years from now; some estimate it be 600 million years (Ward & Brownlee, 2004). This is between 15.7 % and 26 % of the past time of existence of life on Earth. These estimations are close to the result from *life expectancy estimation based on anthropic shadow,* that is, they are around 10 % of past life duration. As we said above, it could be evidence of a strong anthropic pressure. (Note that in the old man example we look not at the whole life duration of 90 years as a base, but only at the difference between his current age and 80; for the Earth, we assume that the beginning of life 3.8 billions of years ago is the beginning of the anthropic stability; before that such events as Late Bombardment had happened. If we look deeper, we could observe that the magnitude of catastrophes has to decline as life became more complex, but some catastrophes may be actually needed to shake biosphere from local minimums.)

Simultaneously, another article (Tyrrell, 2020) modeled climate stability against runaway warming or freezing and found that most simulated planets had very small chance to survive until now. He wrote: “Earth’s climate has remained continuously habitable throughout 3 or 4 billion years. This presents a puzzle (the ‘habitability problem’) because loss of habitability appears to have been more likely. Solar luminosity has increased by 30% over this time, which would, if not counteracted, have caused sterility. Furthermore, Earth’s climate is precariously balanced, potentially able to deteriorate to deep-frozen conditions within as little as 1 million years. Here I present results from a novel simulation in which thousands of planets were assigned randomly generated climate feedbacks. Each planetary set-up was tested to see if it remained habitable over a period of 3 billion years. The conventional view attributes Earth’s extended habitability solely to stabilising mechanisms. The simulation results shown here reveal instead that chance also plays a role in habitability outcomes. Earth’s long-lasting habitability was therefore most likely a contingent rather than an inevitable outcome”.

Tyrrell numerical experiment showed that if a planet reached civilization-level by a pure chance without catastrophic atmospheric instability, it has happened with median around 10 per cent of all simulations (some planets were more successful, but there were only a few of them). In 18 per cent of cases, climate was permanently stable because of lucky initial conditions, which provided stabilization mechanisms, so-called “Gaya”. So, both ways to stability, pure luck and self-regulation, had similar chances. If we look only at the planets which survived by chance, this median survival probability of 10 per cent is equal to anthropic shadow of 1 in 10, which gives 3.8: Log2(10) = 1.18 billion years of future life expectancy, which is surprisingly close to what we already know from Sun’s growing luminosity.

We could also use a frequency of near-misses as a proxy or the real rate of catastrophes. One type of near-misses is especially interesting: it is the situation when the climate was moving to the direction of inhabitability but some unexpected events stopped this progression. There are a least two such events in recent geological history which look at the first glance as such almost catastrophic near-misses: PETM warming 55 millions of years ago and current glaciation in last 3 million years. Two events for around 50 million years give estimation of climate stability of around 25 million years. More about near-misses in Appendix.

## 5.6. The estimation of the frequency of the natural catastrophes via the Fermi paradox

In this section, we show that the best estimate of extinction level natural catastrophes, which is shadowed from us by observation selection, is 1 in 100 million years.

In the observable universe, there are around 1022 stars [ref]. If at least one of them was able to create an advanced civilization, it likely could be able to expand at a speed close to the speed of light, perhaps by sending out small, nanobots-based probes (Armstrong & Sandberg, 2013). Our continued existence and the lack of evidence of any extraterrestrial activity is an argument for that there are no advanced expanding civilizations in our past light cone, that is grabby aliens (Hanson et al., 2021).

The best explanation of the Fermi paradox may be the “Rare Earth” hypothesis (Ward & Brownlee, 2003), namely, that life and civilizations are very rare in the Universe, and there is some Early filter, like abiogenesis (life from non-life). A later Great filter is also possible, but it would have to be *universal*—that is, affect all possible civilizations—which is unlikely (Thrasymachus, 2012). The alternative explanation, that we are the first, is undermined by our not surprisingly early position in time relative to the age of the universe. A new analysis of the Drake equation by Sandberg et al. supports the Rare Earth hypothesis (Sandberg et al., 2017).

The Rare Earth hypothesis means a high level of anthropic selection, as only one from 1022 stars was capable of producing aplanet with intelligent life. This selection consists of the positive fine-tuning of parameters and of the “negative fine-tuning”, that is escaping all types of catastrophes. While it is not clear which one is strongest, we suggest that the forces are comparable.

Many stars in observable galaxies in the universe are very far from Earth, and they appeared very early in the history of the universe and have high red shift, which would prevent even advanced forms of intergalactic travel. Let us assume that only 1020 stars are in the part of the universe to which the Fermi paradox is applicable. Each such star could be regarded, for the purposes of estimation of anthropic pressure, as a failed attempt to create observers.

If negative fine-tuning is a half of all fine-tuning, it is responsible for a square root of the total number of stars, 1010, because probabilities of both types of fine-tuning are multiplied in a way similar to the Drake equation. Thus, the total anthropic shadow’s power for past Earth catastrophes would be 1 to 1010. This means that most of potentially habitable planets die in asteroid impacts, gamma ray bursts and runaway warmings or freezing, and some other catastrophes which humans do not even know about as humanity has never observed them. Only one 1 in 10 billion survives. (This should be affected by universes with larger number of survivors according to SIA, but we don’t know how strong is the effect.)

10 billions is around 33 doublings, and given the Earth’s age of 4 billion years, an event with 50 per cent extinction rate should happen every 120 million years. This estimation is close to the rate of major extinction in the past and suggest that future life expectancy will be the same 120 million years, if we don’t take anthropic fragility into account.

Alternative estimation is the use of remaining habitability of Earth as half-life, that is 1 billion years. In that case, past 4 billions of years is 4 half-lives and total survival chance was 1 to 16. This seems to be lowest level of anthropic shadow, and the real value is higher. There are different estimates of the future habitability of Earth (and they do not take into account anthropic fragility), and Ward gives only 600 million years (Ward & Brownlee, 2004), which means that the past was around 7 half-lives and there was 1 to 128 chance of survival to now.