Back to the Future:

Curing Past Suffering and S-Risks via Indexical Uncertainty,

Acausal Trade and Domination in The Multiverse – or Time Travel

***Alexey Turchin***

Digital Immortality Now

Foundation Science for Life Extension

alexeiturchin@gmail.com

version 3.

**Abstract:** The long unbearable sufferings in the past and agonies experienced in some future timelines in which a malevolent AI could torture people for some idiosyncratic reasons (s-risks) is a significant moral problem. Such events either already happened or will happen in causally disconnected regions of the multiverse and thus it seems unlikely that we can do anything about it. However, at least one pure theoretic way to cure past sufferings exists. If we assume that there is no stable substrate of personal identity and thus a copy equals original, then by creating many copies of the next observer-moment of a person in pain in which he stops suffer, we could create indexical uncertainty in her future location and thus effectively steal her consciousness from her initial location and immediately relieve her sufferings. However, to accomplish this for people who have already died, we need to perform this operation for all possible people thus requiring enormous amounts of computations. Such computation could be performed by the future benevolent AI of Galactic scale. Many such AIs could cooperate acausally by distributing parts of the work between them via quantum randomness. To ensure their success, they need to outnumber all possible evil AIs by orders of magnitude, and thus they need to convert most of the available matter into computronium in all universes where they exist and cooperate acausally across the whole multiverse. Another option for curing past suffering is the use of wormhole time-travel to send a nanobot in the past which will, after a period of secret replication, collect the data about people and secretly upload them when their suffering becomes unbearable.

**Highlights**

* The existence of intense past sufferings a significant moral problem.
* If we create many copies of the next observer-moment of a person in which sufferings end, it creates indexical uncertainty about her future location and effectively saves her from hell.
* To save any past person from sufferings, a future benevolent AI needs to create many possible extensions of each of her observer-moments.
* This requires multiverse-wide acausal cooperation between superintelligences to perform an enormous amount of computations.
* Curing past suffering and resurrection of all the dead is basically the same procedure.
* An alternative way of curing past sufferings is to send a nanorobot into the past via a wormhole – or travelling between everettian branches.

[1. Introduction 3](#_Toc58952561)

[2. Basic algorithm to save a mind from hell 3](#_Toc58952562)

[2.1. The problem of s-risks 3](#_Toc58952563)

[2.2. The nature of indexical uncertainty 4](#_Toc58952564)

[2.3. The basic algorithm of saving minds from “hell” 5](#_Toc58952565)

[2.4. Utilitarian analysis of the trick 6](#_Toc58952566)

[3. Assumptions, limitations and objections 6](#_Toc58952567)

[3.1. Modal realism: everything possible exists 6](#_Toc58952568)

[3.2. Superintelligence is possible, and there are only two relevant types: Evil AIs and Benevolent AIs 7](#_Toc58952569)

[3.3. Evil superintelligences are very rare and everybody knows this 7](#_Toc58952570)

[3.4. Copy-friendly identity theory 8](#_Toc58952571)

[3.5. The measure of existence depends only on the number of copies 9](#_Toc58952572)

[3.6. The duration of sufferings is important 9](#_Toc58952573)

[3.7. Subjective probabilities and anthropics are valid ways of reasoning 10](#_Toc58952574)

[3.8. There are all possible copies in the multiverse, why create more copies? 10](#_Toc58952575)

[4. Computationally efficient ways to save minds 10](#_Toc58952576)

[4.1. Lowering computational cost 10](#_Toc58952577)

[4.2. Lowering the amount of total sufferings 12](#_Toc58952578)

[4.3. Curing past sufferings is technically equal to the resurrection of all the dead and vice versa 12](#_Toc58952579)

[5. Acausal multiverse-wide war against evil 13](#_Toc58952580)

[5.1. Evil AI strikes back 13](#_Toc58952581)

[5.2. How to protect oneself against future s-risks? 14](#_Toc58952582)

[5.3. What could go wrong? 14](#_Toc58952583)

[6. Curing past sufferings via time travel 15](#_Toc58952584)

[6.1. Sending nanobots via wormhole 15](#_Toc58952585)

[6.2. Travelling between branches in the Everett multiverse 16](#_Toc58952586)

[Conclusion 16](#_Toc58952587)

[Literature 16](#_Toc58952588)

# **1. Introduction**

I once heard some awful screams in the middle of the night from a house on fire, which I will never forget. An apartment on the ground floor caught fire, but the windows were barred resulting in a young girl burned alive. I am disturbed by the fact that some people had unimaginably intense pain in the past – what could be done about it?

Moreover, much stronger and potentially infinite sufferings are possible in the future, if a superintelligent AI will have a wrong goal installation or will use human torture as a bargain against other AIs which value humans. This idea is called *s-risks* from “sufferings risk” and is a form of existential risks (x-risks) (Daniel, 2017; Tomasik, 2015). In the last few years scientific literature regarding these risks are explored (Sotala & Gloor, 2017; Torres, 2018). Some people take this possibility so seriously that they think it is better to choose non-existence now than having live long enough and have even a small chance of experiencing infinite torture. Turchin described the idea of acausal multiverse wide cooperation for the particular case of the resurrection of all dead (A. Turchin, 2019b). The idea of indexical uncertainty was also presented before. The [post](https://slatestarcodex.com/2018/04/01/the-hour-i-first-believed/) “The Hour I First Believed” by Scott Alexander (Alexander, 2018) is an introduction to the basic concepts which are used in this study. We believe that by combining all these, we could create an algorithm of saving minds from hell, however, it will work under several assumptions.

The goal of this article is to show that at least theoretically s-risks are curable and thus their possibility should not cease the community from advancement in AI technology. Surely, superintelligent AI will find much better solutions for these problems and will patch up holes which seem unrepairable now. Our latest article on similar content showed that the problem of “end of the universe” is solvable theoretically (Turchin, 2019a).

Section 2 of this study presents the main algorithm in a nutshell while section 3 discusses its assumptions. In section 4, we analyze the ways to make it more computationally efficient. Section 5 describes a more general perspective of acausal war between AIs in the multiverse. Finally, in section 6 we look at other possible ways to cure past sufferings including time travel.

# **2. Basic algorithm to save a mind from hell**

## 2.1. The problem of s-risks

If we are in some form of the multiverse, in which every possible universe exists, s-risk timelines also exist in it, even if they are very improbable. Moreover, these timelines include any actual living person, including the reader. This thought is disturbing. This idea initially appeared in the science fiction story “I have no mouth and I must scream” (Ellison, 1967).

## 2.2. The nature of indexical uncertainty

The main idea, presented below, is based on the notion of “indexical uncertainty”. This is a situation when an observer doesn’t know where he is located if numerous copies of him are created. For example, if 100 copies of a person are created and all are put in identical boxes, of which all except one is red from outside, and one is green then according to Self-Sampling Assumption in anthropic (Bostrom, 2013), one should reason that the copies are randomly selected from all of them and thus he has 99 per cent chances to observe that he will be in the red box.

Therefore, if some agent creates many copies of the person in location X, he could attract the probability mass that the person should expect to find himself in the location X, even if the person was originally in the location Y.

Indexical uncertainty also works for *my future copies*. If I expect that several my copies will be created in the next moment, I can expect that my probability to be one of them is proportional to the share of this type of copies relative to the total number of copies. It is based on some assumptions about personal identity which will be discussed later. Indexical uncertainty is used, for example, to derive Born rule in many world interpretation of quantum mechanics by Sean Carrol (Carroll & Sebens, 2014).

It is a known idea to use indexical uncertainty for the attack in AI boxing experiment. An AI says to an operator that it has created a million his copies of the operator, and it will torture all these copies if the operator will not let him out. Now the operator is uncertain if he is real or simulated and may decide to escape the risk and let the AI out.

This way of reasoning about indexical uncertainty is based on the *Self-Sampling Assumption* (SSA). There are other ways to reason about anthropics. *Self-Indication Assumption* (SIA) doesn’t look like as a contradicting method of reasoning here (Bostrom, 2013). But the *Updateless Decision Theory* (UDT) and its variants may give different results (Armstrong, 2011): oversimplifying, in this theory one should care not about probabilities of his individual copies, but about global consequences of a given strategy in terms of achievable utility. In our case, an agent under intensive sufferings has very small choice about future actions and very strong preference to stop his own sufferings (he doesn’t have mental energy to be a universal altruist); also in most cases, he even doesn’t know that he will be saved in the future, but he could hope according to his belief system. If he uses *meta-level uncertainty* about the validity of different philosophical concepts (Turchin, 2018a), he would still give 10-50 per cent credence that the whole trick will work, and it is a good chance to escape sufferings (many cancer therapies have lower success rates).

## 2.3. The basic algorithm of saving minds from “hell”

S-risks prevention could be realized via the following "salvation algorithm":

Let S(t) be an *observer-moment* of an observer *S* who is experiencing intensive suffering at time step *t*.

In the [logical time](https://www.lesswrong.com/posts/dKAJqBDZRMMsaaYo5/in-logical-time-all-games-are-iterated-games) (Demski, 2018) the sequence of the “salvation algorithm” is as following:

1. S(t) is suffering in some Evil AI's simulation in some causally disconnected timeline.

2. A benevolent superintelligence creates 1000 copies of S’(t+1) observer-moments[[1]](#footnote-1), which are next observer-moments for S(t) but on which suffering has stopped or significantly relieved.

3. Now, each S(t) is uncertain where will be its next observer-moment: in the evil simulation, or in a Benevolent AI’s simulation – but, using the self-sampling assumption, S(t) concludes that with the probability of 0.999% it will be located in the Benevolent AI’s simulation. (Note that because we assume the causal connection between observer-moments is not important, it is equivalent as moving to the Benevolent simulation).

4. A Benevolent AI creates 1000 continuing observer-moments S’(t+2) where remaining sufferings gradually but quickly decline, and each such observer-moment is a continuation of S(t+1) observer-moment. The pain declines “gradually” in order to preserve identity as a consistent chain of moments. For human-level pain, it will be an *instant relief*, and only some memories about having this pain will remain. But strong memories about pain themselves could be traumatic in case of PTSD, so the next step in relieving pain is needed to mitigate the impact of the memories.

5. The Benevolent AI creates a separate timeline for each S’ (t+1) observer-moment, which looks like S’ (t+2)….S’ (t+n), a series of experiences wherein the observer becomes happier and happier.

6. The Benevolent AI merges some of the timelines to make the computations simpler.

7. The Evil AI creates a new suffering moment, S(t+1), in which the suffering continues.

8. The benevolent AI performs all steps 1-7 for this next S(t+1) observer-moment of sufferings.

Thus, from the point of view of any suffering-observer-moment S(t), the future of the observer-moment is dominated by timelines where he is saved by a Benevolent AI and will spend eternity in paradise.

## 2.4. Utilitarian analysis of the trick

This “salvation algorithm” may work not only for fighting Evil AI but for addressing any type of past suffering. For animal lovers, an additional benefit is that this approach will also work to undo all past animal suffering, including that of the dinosaurs or aliens.

In some sense, we just dissolve the large number of suffering-observer-moments with astronomically larger number of happy-observer-moments (for each type of observer), and thus any given observer has very small chances to be in the situation of s-risk and will escape it almost immediately.

Note that he existence of my luckier copies seems inevitable. There are branches of the multiverse where I am richer, healthier and my friends and relatives are still alive. If benevolent AI will appear, it will also create infinitely many incredibly happy people. The main difference is that in this setup I can rationally expect that my next observer-moment will be such world. The main difference of this setup is not the existence of lucky copies, but a distribution of measure in their favor.

# **3. Assumptions, limitations and objections**

Eternal torture by Evil AI in some unlikely future (s-risks) becomes possible under several assumptions, but almost the same set of assumption makes the salvation from these timelines via indexical uncertainty possible.

## 3.1. Modal realism: everything possible exists

If everything possible exists (Tegmark, 2009; Turchin, 2018b), there will be branches of multiverse where evil AI which wants to torture people will appear for sure. No matter, if it is 1 percent of all possible variants of future or 0.0001 per cent, it is disturbing to know that human beings and even our copies will be tortured. AI-empowered torture may be worse than anything human can imagine, because the digitalized mind is almost immortal, so there is no escape from death. As the evil AI could upgrade human mind to make it capable to feel even more pain than normal human mind, and even continue to upgrade it eternally, its sufferings will only constantly increase. A perspective of eternal and constantly increasing torture is so bad that many people regard it even worse than total human extinction, even if it has the slightest probability. Modal realism makes s-risks inevitable somewhere.

However, if everything possible exists, very large computations could be performed by distributing parts of them between different parts of the multiverse, thus making sure that a counterpart exists who will perform their part of computations. Thus, a large project of saving minds from hell via indexical uncertainty is possible.

Modal realism also defines which minds are needed to be saved: it is all possible minds. For any person, there will be a universe where an advance AI will torture this person. Hence every possible person needs to be saved at each moment of her existence. However, any person could also die at any moment, so this task is similar to the resurrection of all possible minds.

## 3.2. Superintelligence is possible, and there are only two relevant types: Evil AIs and Benevolent AIs

S-risks become possible only if a digital, eternal mind with superhuman capabilities which can have stable (but negative) goals system is possible. The same type of mind is needed to cure s-risks, as we need to perform enormous amounts of computations to create and maintain copies.

AIs with any possible goals are feasible (Bostrom, 2014), like Paperclip maximizers, but they will “just” kill all people for atoms, without deliberately causing sufferings. Other AI may just not care about humans. From the sufferings point of view, only two types of AIs are important: Benevolent AIs, which are actually interested in human wellbeing, and Evil AI, which has the goal or sub goal of torturing humans.

## 3.3. Evil superintelligences are very rare and everybody knows this

We also assume that Evil AIs are very rare, as they appear because of some very unlikely mistakes, like replacing “+” with “­–” in the utility function of some benevolent AI. Some Paper clippers may have torture as sub goal if they think that it gives them barging point against benevolent AI. E.g. “I will not torture people in the simulation if you will give me your resources.” However, this most likely works if both AIs are in one physical world. It is less likely that such blackmail will work acausally. But *Roko Basilisk* thought experiment (in which future AI will punish humans which will not help its existence) works acausally (Auerbach, 2014), except that it hasn’t actually worked: nobody has fallen to this blackmail, as far as we know.

We also assumed that creation of Benevolent AI is a solvable task and it will be solved in the largest group of the worlds (“largest part” compared to the share of evil AIs worlds: the AIs which painlessly exterminate humanity are not counted here). In other words, Benevolent AIs could have access to a million times more computational resources but are located in different branches of the multiverse (which is not necessarily a quantum multiverse but maybe an inflationary one, or of some other type (Knobe et al., 2006)).

To win in the acausal game, one needs to successfully predict what is happening in other branches of the multiverse and make sure that everyone else has the same knowledge. This makes cooperation possible (see also a recent article where shared knowledge is needed for SETI contact, which explains it based on game theory (Kerins, 2020)).

One argument for the rareness of s-risks is that we are not in such a situation. It is based on the Copernican mediocrity principle, which claims that humans (and me-now in particular) are typical observers in a typical world. As most people are not in a situation of intense sufferings now, we could exclude the idea that Evil AIs are dominating the whole measure mass of the observers in the universe. Thus, we can exclude the idea that 99 percent of observers are in hell, barring 50 percent of them are in hell. General counterarguments against Doomsday argument could be applied here, like non-randomness of the question (Turchin, 2018a).

## 3.4. Copy-friendly identity theory

Copy-friendly identity theory states that only similarity of observer-moments counts for identity, not the “continuity of consciousness” (Bamford & Danaher, 2017). If this is not true, *hostile resurrection* is impossible and we are mostly protected from s-risks, as suicide following by eternal death becomes a possible option. Our previous work shows that either some immaterial soul exists, or copy equals original. If an immaterial soul does exist, then God is also likely to exist and such God could be benevolent. Thus, it will likely be able to stop s-risks (Turchin, 2020).

To save minds or resurrect them, there should be no soul, and there is no causal continuity of consciousness as a substrate of identity. Each observer-moment is connected with another one only because it has memories about it or anticipations of the future.

But some form of identity must exist for the possibility of s-risks. It looks like s-risks is not a problem for *open, closed or empty individualism*. If it is empty, it is not me who will suffer; if it is closed, eternal suicide is possible and hostile resurrection is not possible. In case of open individualism, I am everybody, but the order of the observer-moments doesn't matter, so they could be deluded by happy observer-moments in the future.

The remaining is the identity theory based on memory: if I am in the state O1, the next observer-moment O2 is the one which remembers O1 as previous observer-moment: memory(O2)=O1. This is close to how we live in real life, where we use our memories of the past to calculate our identity. The identity-based-on-memory allows many next observer-moments for any previous observer-moment and doesn’t require a real causal connection between observer-moments.

## 3.5. The measure of existence depends only on the number of copies

*The Measure of existence is defined* over different branches in the quantum multiverse as a square of the amplitude of the branches’ wave function, proportional to its probability. It could also be defined asa subjective probability that one will find oneself in one of the states of mind (Vaidman, 2018). Typically, it is assumed that if two equal copies of a person are created, then he has equal chances to be each of them.

However, it is not all that simple. For example, if copies are located in different branches of quantum multiverse, one should multiply the number of copies on the probability of each branch. Also, if the copies run on different substrates, like biological or digital, it could also affect the subjective probabilities in unknown ways. Yudkowsky has demonstrated this by the idea of *ebborians*: the minds, which have a thickness in the third dimension and could be split differently thus affecting subjective probability (Yudkowsky, 2008c). Extrapolation of this idea on minds in general offers that *minds which use more energy for computations are more probable* (as they could be virtually split on lower energy brains running in parallel).

To save a mind via indexical uncertainty is the same as to get access to the highest measure of the given instances of the mind: the highest measure means the biggest subjective probability for that mind to be present here. If it turns out that the number of copies is not enough to “win measure”, some other methods could be used, like increasing energy of computation or increasing the number of small variations which creates the set of minds of the given type, or some manipulations with qualia (as having qualia is needed to actually exist, and maybe some “brighter” qualia correspond to a higher measure of existence). We could assume that superintelligent AI will solve the measure problem and will find the ways to get access to the highest measure if it has enough resources.

Note that we assume here that most minds in s-risks situations are not biological, but are digital-mind-models inside computer simulations, and this doesn’t have any subjective difference or affect the quantity of experience (measure). But past minds were biological, and we have to assume that each digital mind has the same amount of measure as biological, and all the same qualia.

## 3.6. The duration of sufferings is important

One more assumption is about the nature of sufferings. It says that sufferings are bad not only based on their *intensity* but also on their *duration* (and impossibility to escape them). This may contradict pure utilitarian view where pain is additive (see e.g. discussion here (Yudkowsky, 2008b)). But the common sense about pain takes into account both intensity and duration.

Another sufferings-related assumption is that there is a level of “unbearable pain” which is really bad, but the pain below this level is normal and could be ignored. We could formally define *unbearable pain* as such a combination of intensity, duration and *expected endlessness of pain* to which a person would rationally choose suicide. It seems irrational to commit suicide because of tooth pain for one night (as did a character of Jack London short story), but a cancer patient or a person in fire may want this.

Also, if quantum immortality is true, suicide is not an escape from pain, but could make it even stronger. Moreover, quantum immortality could be itself a source of s-risk, as any person will inevitably age and will get terminal illnesses, but will not be able to die, as there will always be branches of the multiverse where he is still alive. The only escape is again benevolent AI in the future, even in a small fraction of future timelines.

## 3.7. Subjective probabilities and anthropics are valid ways of reasoning

The whole trick is based on the Self-Sampling Assumption in anthropics. Many scientists deny anthropics, or suggest different interpretations, like *Anthropic decision theory* (Armstrong, 2011), in which not subjective probabilities, but most optimal decisions are calculated.

## 3.8. There are all possible copies in the multiverse, why create more copies?

These copies are not your next observer-moments or they have very low measure and thus there is a very small probability that you find yourself in them in the next moment. For example, for every cancer patient, there is a universe where he will be informed that he is cured in the next moment, but the share of such branches is so small that they are very unlikely and I unlikely to find myself in that moment. Note that here is assumes that the probability of finding oneself in some specific next observer-moment depends only on its measure, but not on "continuity of consciousness".

# **4. Computationally efficient ways to save minds**

## 4.1. Lowering computational cost

The problem with the above-suggested model is its computational cost: for any suffering observer-moment, 1000 full lives must be simulated. Several ways to lower such costs can be imagined:

**Patch 1.** The size of the observable universe is limited and thus Evil AI and any particular Benevolent AI will have (in the end) similar computational resources. But the number of universes with Benevolent AI is assumed to be larger. In that case, different Benevolent AIs may distribute parts of the task between each other using randomness, where each of them randomly chooses the part of work it will do, and after a large number of attempts all (or almost all) work is done.

Generating random minds also solves the problem that any Benevolent AI will not know which observers are suffering, and thus it will have to save all possible suffering observers (thus, it must counterfactually model the existence of all possible Evil AIs—or maybe not, perhaps only possible suffering observers).

Evil AI also could use random mind generation for torturing. At step 1 it creates a random file using a quantum randomness generator. In Step 2, it checks if this file is a description of a mind. E.g. using random bits as weight in a neural net. In step 3, it explores this mind and creates the most displeasing condition for it. If Everett multiverse is true, this is enough to torture every possible mind in some of the branches of the multiverse. By creating many copies of this machine, the measure of tortured minds could be increased. Here “many” means tiling the whole visible universe with such machines and cooperating with other universes. However, Benevolent AI may use almost the same machine to resurrect all possible minds and save suffering observer-moments. It wins if it has a larger number of such machines.

**Patch 2.** Not every sufferings-observer-moment should be counted as an observer needing to be saved. First, there is a threshold of unbearable suffering, discussed above, something which a normal person could feel only in a moment of intense agony, like if the whole body is on fire. Some moments of intense pain are normal, and those experiencing them do not require “saving”, e.g. the pain of childbirth. But Evil AI may create much stronger sufferings.

**Patch 3.** If the suffering could be bearable for a short period, for example, for 2 seconds, not every observer-moment needs to be saved, but only those that are separated by some time-distance, after which it becomes unbearable.

**Patch 4.** The number of new moments doesn’t need to be large. Even if for every suffering observer-moment just one copy is created, after 10 steps of internal time the subjective chances that the observer is still in the original timeline will be 1 in 1024. The main constraint here is that the efforts toward salvation should be stronger than any possible attempts of the Evil AI to “steal” observers back from the Benevolent AI.

**Patch 5.** After minds are saved from suffering, they could merge and after this, the amount of computation used to model them may be lowered. This reminds of the “Anthropic trilemma” by Yudkowsky (Yudkowsky, 2009), where many copies are created to change the subjective probability of experiencing the winning in a lottery, but later, all these copies are merged back. This, however, creates several paradoxes about anthropic reasoning such as either magic works, or probabilities behave inconsistently, or personal identity is arbitrary, or paradoxical claim that “to deny that increasing the number of physical copies increases the weight of an experience”. Our solution of s-risks is akin to winning in a lottery in anthropic trilemma and is requiring that “anthropic magic” to work.

## 4.2. Lowering the amount of total sufferings

**Patch 1**: The (initial version of) salvation algorithm required creating observer-moments which are copies of suffering observer-moments to insert indexical uncertainty. This initial version follows from the idea of indexical uncertainty and was discussed in the earlier versions of this paper – but was correctly rejected by peers. Here, paradoxically, even if no evil AI exists, they will be modeled and all their victims will be created only to be saved. Even if we patch utilitarianism to regard exact copies as the same moral object with the same value, still some unease remains.

To counter this objection, the step 4 of the algorithm, where 1000 copies of the moment of suffering are created, was patched. Instead of creating indexical uncertainty of a current observer-moment S(t), we will create indexical uncertainty for the next observer-moment S(t+1), which will have memories of having sufferings at the moment t, but no sufferings now. We will do it by creating a large number of copies of the mind S, starting not from the moment t, but from the moment t+1.

Imagine a person who is dying from cancer at 12.00 AM. At 12.00.01 we create 1000 copies of her which are informed of being cured. If copy equals original, then this person may expect that with 0.999 per cent probability her next observer-moment will be in a healthy cured body.

**Patch 2:** Another way to escape the increase in the number of total suffering is to use indexical uncertainty attack not to the moment of sufferings S(t) but to the previous moment S(t-1), when sufferings didn’t start yet at all. If the observer-moment S(t-1) is diluted by creating many copies in paradise, then the subjective probability to appear in hellish S(t) is very low. This, however, will not work for a long chain of sufferings, as all previous moments are also hellish. Hence, it helps against falling in hell, but less effective from saving one from hell.

## 4.3. Curing past sufferings is technically equal to the resurrection of all the dead and vice versa

Tippler in his Omega point (Tipler, 1997) suggested to resurrect all possible minds using enormous computation resources of collapsing universe and Almond (Almond, 2006) suggested more computationally efficient way to do it via quantum random mind generator. I suggested even more effective way to do it, by adding cooperation between different universes and use of quantum immortality data (Turchin, 2019b).

Any person who is now suffering could be dead in the next moment, and he will die in some future timelines and will survive in another one. Moreover, for any person who is now dying, there is a timeline in which he will not die in the next observer-moment – this is the idea of *quantum or big world immortality* in a nutshell (Turchin, 2018b).

Thus, resurrecting of any possible person is equal to saving from sufferings any possible person. The main difference here is the *measure*. In our original setup of saving minds from hell, we created 1000s times more copies for every suffering copy to get the measure domination – but in the case of the resurrection, we don’t need *measure* domination, as the *measure* of the dying person quickly declines.

However, imagine that we create just one copy of a person every 1 second. He can estimate his subjective chances to find himself as an original in the next second as 0.5. After repeating coping this for 10 times in 10 seconds, he will have only 1 in 1024 chance to find himself as an original. In other words, *repeating resurrection creates measure domination,* which is needed for curing past sufferings.

Thus, there will be no two different procedures for curing past sufferings and for the resurrection of the dead that will create two different groups of minds which need to be reconciled. There will be just one technology.

# **5. Acausal multiverse-wide war against evil**

## 5.1. Evil AI strikes back

Evil AI or a coalition of such AIs may try to fight back and steal a mind back from paradise by creating even larger numbers of copies. There are two main reasons why this is unlikely to happen:

1. Evil AIs can’t cooperate (maybe because they are evil for different reasons), OR
2. All Evil AIs have total measure that are millions of times smaller than the measure of collaborating Benevolent AIs. Even if an Evil AI steals some observers from paradise, the Benevolent AI could regain these observers via its salvation algorithm in just nearly-immediately.

However, Evil AI may steal just a few observers chosen randomly, so benevolent AI will never guess who ends in hell but will know that someone will. It is a difficult task as Evil AIs in different branches can’t send each other data about the observer which will be stolen without benevolent AI being able to predict this. In other words, Evil AIs can’t cooperate *acausally and secretly*, as any acausal deal, by its nature, is open, and Benevolent AIs will know who is chosen to go to hell. But if an Evil AI spends all available resources of one observable universe for the torture of just one randomly selected observer, it could succeed, as the total measure of this observer will be very large.

Obviously, the coalition of benevolent AIs may have its answer to this, but this is out of the scope of this study. We can’t predict the next steps of advanced AIs. However, we could describe the general situation as *acausal war* between AI coalitions.

To dominate in the multiverse, a coalition of AIs needs to control as many observer-moments as possible and thus any conscious being or AI may be sure that he is in a simulation created by such AIs. This means that all AIs will have strong incentive to create simulations, and also to increase the measure of minds in them, by creating “anthropic monster” (analogue of “utility monster”) thus proving the Simulation argument (Bostrom, 2003).

A novel [*Surface Detail*](https://en.wikipedia.org/wiki/Surface_Detail) by Iain M. Banks describes some kind of war around artificial hells (Banks, 2010).

## 5.2. How to protect oneself against future s-risks?

Destroying one’s human digital footprint will not help to protect against hostile resurrection (some people have suggested this as an argument against *indirect digital immortality*), if Evil AI recreates all possible beings. Instead one should invest in creating Benevolent AI, interested in the resurrection of the dead and saving suffering observer-moments.

## 5.3. What could go wrong?

The measures discussed here assumes a very ambitious value system which could affect the wellbeing of all possible minds in the multiverse. As Yudkowsky wrote, we should very careful then we do something very close to the s-risks in the space of possible solutions (Yudkowsky, 2014). Any mistake could actually cause s-risks, not prevent them or exacerbate them. One concern is that creating copies of suffering mind actually increases the total sufferings in the universe.

Also, a Benevolent AI may be only a “benevolent” AI and does a mistake similar to tiling the whole universe with photos of smiling faces instead of creating happy people (Yudkowsky, 2008a). So, we should not enforce AI to implement this particular solution of s-risks’ prevention before it reaches superintelligence level and will have a proven benevolence – and will able to decide is it actually good or not.

Fighting possible Evil AIs is also a large opportunity cost as to win such war one has to use almost all available resources of the whole observable universe. This may cause *a-risks*: that is, risks to the existence of other alien civilizations which will not appear because of earthlings’ colonization.

Another opportunity cost is that we can create many much happier people *de novo*, or do some other very interesting things which we can’t imagine now.

Benevolent AIs may also lose this war.

# **6. Curing past sufferings via time travel**

## 6.1. Sending nanobots via wormhole

If time travel was possible, we could return to the past and save people from horrific sufferings. Similarly, the possibility of traveling between branches of the multiverse could help to directly confront Evil AIs. However, the only known possible way of time-travel at present requires creation of wormholes, which will distort space-time near it as it is as strong as a black hole (Morris et al., 1988). The appearing of such a wormhole near the Earth would destroy the planet and moving of large objects via the wormhole is unlikely because of strong gravitational force in it.

Moreover, the location of the exit of a natural wormhole is likely near the beginning of the universe (as it is the place when a lot of strange high-energetic events happened), and it’s another exit may be very far, probably in another galaxy, due to the rarity of such objects.

As the wormhole’s *aperture* is likely to be very thin, it will allow passing of only very small things. Thus, we could send just one nanobot back in the past. This nanobot needs to be capable of self-replication and have AI onboard (Drexler, 1986). After occurring in the past, it will self-replicate and eventually reach the Earth as von Neumann probes (Armstrong & Sandberg, 2013). Nevertheless, it should do this in a hidden way in order to escape “grandfather paradox” (precluding its own existence via changes in the past (Lewis, 1976)). Such nanobots will be able to secretly live in the brains of people and record their activity and turn off suffering if it becomes unbearable.

Such nanobots also could directly prevent creation and activity of Evil AIs, if they will be created on Earth (or by other civilizations).

The “butterfly effect” (small changes of the future because of time travel to the past, which is a mild form of grandfather paradox) seems to be inevitable, but it will end in a strange attractor of infinitely many loops. In other words, we don’t save our past, but the past of some other similar world. However, it is very likely that our past was saved by some other time travelers, very similar to us. It reminds the acausal trade between universes discussed above.

## 6.2. Travelling between branches in the Everett multiverse

Another way to cure s-risks is to find the way to travel between branches in Everett multiverse and invade those branches there Evil AIs have appeared. Obviously, it is a bidirectional way, so there should be protection from Evils AIs invasion.

Surprisingly, one possible instrument for such travel may be similar to creating indexical uncertainty. One possible way to travel between branches is the use of “many pasts” theory, which is the extension of the multiverse theory not only to the future but to the pasts. The theory postulates that all past histories resulting in curing conditions are equally real. To travel, we prepare a traveler in the condition which could both appear in our timeline and in the destination timeline, and then the traveler prepares a measurement in which timeline he is located.

Roko Basilisk thought experiment is an example of “acausal channeling” from the one possible future timeline to the past (of many more timelines), but here is used a decision theory trick, not indexical uncertainty; actually, here a semi-evil-AI moves into the past.

Other ways to travel between timelines may be more physical and use *residual coherence* between separated branches, but it is closer to science fiction now.

# **Conclusion**

In this article, we presented two independent ideas about curing past sufferings and s-risks: via indexical uncertainty and via time travel. Both are based on a series of assumptions. It shows us that the problem is tractable, though the future seems to be open to other ideas. As the problem is tractable, there is no need to commit suicide or stop technological progress because of the possibility of s-risks. However, we should include the idea of curing past sufferings and preventing s-risks into future possible goal function of Global AI which irreversibly become a world government and is called Singleton (Bostrom, 2006).

# **Literature**

Alexander, S. (2018). The Hour I First Believed. *Slate Star Codex*. https://slatestarcodex.com/2018/04/01/the-hour-i-first-believed/

Almond, P. (2006). *Many-Worlds Assisted Mind Uploading: A Thought Experiment*. https://web.archive.org/web/20110513092111/http://www.paul-almond.com/ManyWorldsAssistedMindUploading.htm

Armstrong, S. (2011). Anthropic decision theory. *ArXiv Preprint ArXiv:1110.6437*.

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

Auerbach, D. (2014). *The Most Terrifying Thought Experiment of All Time*. Slate. http://www.slate.com/articles/technology/bitwise/2014/07/roko\_s\_basilisk\_the\_most\_terrifying\_thought\_experiment\_of\_all\_time.html

Bamford, S., & Danaher, J. (2017). Transfer of Personality to a Synthetic Human ('Mind Uploading’) and the Social Construction of Identity. *Journal of Consciousness Studies*, *24*(11–12), 6–30.

Banks, I. M. (2010). *Surface detail*. Hachette UK.

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

Bostrom, N. (2006). What is a Singleton? *Linguistic and Philosophical Investigations*, *2006*(5).

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

Bostrom, N. (2014). *Superintelligence*. Oxford University Press.

Carroll, S. M., & Sebens, C. T. (2014). Many worlds, the born rule, and self-locating uncertainty. In *Quantum theory: A two-time success story* (pp. 157–169). Springer.

Daniel, M. (2017). *S-risks: Why they are the worst existential risks, and how to prevent them (EAG Boston 2017)*. https://foundational-research.org/s-risks-talk-eag-boston-2017/

Demski, A. (2018). *In Logical Time, All Games are Iterated Games—LessWrong*. https://www.lesswrong.com/posts/dKAJqBDZRMMsaaYo5/in-logical-time-all-games-are-iterated-games

Drexler. (1986). *E.: Engines of Creation*. Anchor Press.

Ellison, H. (1967). *I Have No Mouth, and I Must Scream*. Galaxy Publishing Corp.

Kerins, E. (2020). Mutual detectability: A targeted SETI strategy that avoids the SETI Paradox. *ArXiv Preprint ArXiv:2010.04089*.

Knobe, J., Olum, K. D., & Vilenkin, A. (2006). Philosophical implications of inflationary cosmology. *The British Journal for the Philosophy of Science*, *57*(1), 47–67.

Lewis, D. (1976). The paradoxes of time travel. *American Philosophical Quarterly*, *13*(2), 145–152.

Morris, M. S., Thorne, K. S., & Yurtsever, U. (1988). Wormholes, time machines, and the weak energy condition. *Physical Review Letters*, *61*(13), 1446.

Sotala, K., & Gloor, L. (2017). Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, *41*(4).

Tegmark, M. (2009). The multiverse hierarchy. *ArXiv Preprint ArXiv:0905.1283*.

Tipler, F. J. (1997). *The physics of immortality: Modern cosmology, God, and the resurrection*.

Tomasik, B. (2015, April 10). Risks of Astronomical Future Suffering. *Center on Long-Term Risk*. https://longtermrisk.org/risks-of-astronomical-future-suffering/

Torres, P. (2018). Space Colonization and Suffering Risks: Reassessing the “Maxipok Rule.” *Futures*.

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. (2019a). *How to survive the end of the universe*. https://philarchive.org/rec/TURHTS-2

Turchin, A. (2019b). *You only live twice: Resurrection of the dead in computer simulations*.

Turchin, Alexey. (2020). *Immortality and The Hard Problem of Identity: Either My Copy Will Be Me – or Soul Exists*.

Vaidman, L. (2018). Many-Worlds Interpretation of Quantum Mechanics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2018/entries/qm-manyworlds/

Yudkowsky, E. (2008a). *Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks* (M. M. Cirkovic & N. Bostrom, Eds.). Oxford University Press: Oxford, UK.

Yudkowsky, E. (2008b). Torture vs. Dust Specks—LessWrong. *LessWrong*. https://www.lesswrong.com/posts/3wYTFWY3LKQCnAptN/torture-vs-dust-specks

Yudkowsky, E. (2008c). Where Physics Meets Experience—Less Wrong. *LessWrong*. http://lesswrong.com/lw/ps/where\_physics\_meets\_experience/

Yudkowsky, E. (2009). *The anthropic trilemma*. http://lesswrong.com/lw/19d/the\_anthropic\_trilemma/

Yudkowsky, E. (2014). *Separation from hyperexistential risk*. https://arbital.com/p/hyperexistential\_separation/

1. using the randomness generator and resurrection model described in the article (A. Turchin, 2019b). In a nutshell, the benevolent AI creates all possible random minds via random code generator. More about it later. [↑](#footnote-ref-1)