**Approaches to the Prevention of Global Catastrophic Risks**

*Human Prospect, 2018, Volume 7 N2 pp.53-65*

*Alexey Turchin*

Digital Immortality Now

Foundation Science for Life Extension

alexeiturchin@gmail.com

**ABSTRACT**: Many global catastrophic and existential risks (X-risks) threaten the existence of humankind. There are also many ideas for their prevention, but the meta-problem is that these ideas are not structured. This lack of structure means it is not easy to choose the right plan(s) or to implement them in the correct order. I suggest using a “Plan A, Plan B” model, which has shown its effectiveness in planning actions in unpredictable environments. In this approach, Plan B is a backup option, implemented if Plan A fails. In the case of global risks, Plan A is intended to prevent a catastrophe and Plan B to survive it, if it is not avoided. Each plan has similar stages: analysis, planning, funding, low-level realization and high-level realization. Two variables—plans and stages—provide an effective basis for classification of all possible X-risks prevention methods in the form of a two-dimensional map, allowing the choice of optimal paths for human survival. I have created a framework for estimating the utility of various prevention methods based on their probability of success, the chances that they will be realized in time, their opportunity cost, and their risk. I also distinguish between top-down and bottom-up approaches*.*

Keywords: existential risks, survival, global warming, preparedness, refuges

[1. Introduction 1](#_Toc530312855)

[2. Principles of classification of prevention plans 3](#_Toc530312856)

[2.1. Plan A and Plan B model. Logical steps in catastrophic prevention as a base for classification 3](#_Toc530312857)

[2.2. Time stages of the realization of a plan 4](#_Toc530312858)

[3. Plans A and B for global risks prevention 5](#_Toc530312859)

[3.1. Overview of global risk prevention plans 5](#_Toc530312860)

[3.2. Plan A: prevent the catastrophe 6](#_Toc530312861)

[3.3. Plan B: survive the catastrophe 8](#_Toc530312862)

[3.4. Map of X-risk prevention plans 9](#_Toc530312863)

[4. Navigating prevention plans’ space 10](#_Toc530312864)

[4.1. Utility and probability of success for different plans 10](#_Toc530312865)

[4.2. Individual plan’s efficiency 12](#_Toc530312866)

[4.3. Strategies to realize prevention plans 12](#_Toc530312867)

[4.4. Musk’s Plan A and Plan B 13](#_Toc530312868)

[4.5. Combining plans in the most safe and cost-effective way 13](#_Toc530312869)

[5. Conclusion 14](#_Toc530312870)

[References 15](#_Toc530312871)

# 1. Introduction

This article is written from the point of view of a transhumanist. The transhumanist world view is that the future of the world will be shaped by several super technologies - nanotech, biotech and strong artificial intelligence - and that these technologies are developing at an exponential rate, so they will reach maturity at some point in the 21st century[1]. The maturation of these technologies facilitates massive new possibilities, like radical life extension, but also new dangers, like complete human extinction. This article is devoted to examining approaches to prevent such dangers.

There are many important things in the future of humanity, but the main question is if humanity will have a future at all. The risks of nuclear war, irreversible global warming and resource depletion have been recognized starting in the 20th century. In the 21st century, it becomes clear that rapid progress in new technologies like nanotech[2], synthetic biology[3] and artificial intelligence (AI) could create even more serious risks, which could end life on Earth. Such risks have been called “existential risks”[4].

The idea of existential risks (X-risks) appeared in the wider context of transhumanism, which aims to solve humanity’s biggest problems first. Nick Bostrom originated the idea that prevention of existential risks should be the primary goal of humanity[5], as all other faults could be repaired, but human extinction is an irreversible event[4]. He also showed that humanity has more than ten extinction risks in the 21st century, most of which are connected with new technologies. Martin Rees estimated the extinction probability for humanity in the 21st century at 50 percent[6].

While large lists of risks have been made, the list of prevention methods is smaller, and in actuality just one comprehensive prevention plan is known. That plan was suggested by Elon Musk; to create safe AI by using OpenAI – a collaboration of different AI teams based on openness and on creation of upgraded human beings via neural implants in his Neuralink project[7] - and to prevent resource depletion by switching to a renewable economy based on solar energy and electric cars. His backup plan is to create a colony on Mars as a survival option[8].

Other people have suggested various ideas of X-risk prevention. Eliezer Yudkowsky[9] suggested creating “Friendly AI,” a beneficial superintelligence that would be able to stop any other existential risk. David Brin[10] spoke about a “transparent society,” where everybody can monitor the behavior of everyone else. Bostrom[4] wrote about “differential technological development,” where beneficial technologies will be implemented first. He also underlined the need for a “Singleton,” a single upper level decision-making agency necessary to solve problems of global risks. Posner[11] advocated for legal and economic approaches to existential risks, like carbon trade.

Recently, several articles on the topic of how to survive a global catastrophe in refuge accommodations have been published, including articles about food security by Denkenberger[12] as well as works by Jebary[13], Baum[14], Torres[15] and Beckstead[16] about various types of refuges. Turchin created a map of possible refuges[17] and wrote articles about nuclear submarines as refuges[18] and survival on islands[19].

There are two possible approaches to X-risks prevention: general top-down approaches and bottom-up prevention of specific risks. General approaches prevent a large spectrum of risks, such as via creation of friendly artificial intelligence, survival bunkers, or technology relinquishment. General approaches work from top to bottom, and may even help to prevent unknown risks. In this article, I will discuss such general prevention measures.

A bottom-up approach to global risks could be very useful in specific, narrow fields, like creating flu vaccine stockpiles or asteroid deflection. But we can never know if all possible specific risks are covered and in the correct proportion. Much research has been conducted on prevention of specific risks, like global warming via geoengineering[20], nanotechnology risks via nanoshield[2], and asteroid defense[21].

The objective of this article is to explore general top-down approaches to X-risks prevention and present their structural classification. In Section 2, principles of classification of the prevention plans will be explored, Section 3 will present Plan A and Plan B approaches, and in Section 4 how to choose most cost-effective plans from all possible x-risks prevention plans will be discussed.

# 2. Principles of classification of prevention plans

## 2.1. Plan A and Plan B model. Logical steps in catastrophic prevention as a base for classification

There is a well-known established way of classifying plans of action, the “Plan A, Plan B” approach used, for example, in nuclear safety engineering[22]. If Plan A fails, Plan B is initiated, increasing the total probability of success.

There are two possible approaches to Plan A: one is to choose the most effective plan, while another is to choose the plan that is closer to the causal roots of the problem.

The “Plan A, Plan B” model can be viewed as different levels of defense against a given risk. For example, in the case of nuclear war, Plan A is to prevent it, Plan B is to protect a country using a missile shield, and Plan C is to survive the war using civil refuges[23].

I think that Plan A for global risks prevention should be the one which is closer to the causal roots of the problem, as it allows a chance to try both Plans A and B. For example, Plan A for fighting global warming is reducing carbon emissions, and Plan B is geoengineering by harvesting carbon from the atmosphere[24]. I could continue this principle by adding Plan C, and so forth. In the case of global warming, Plan C is emergency measures, like dimming the atmosphere with aerosols.

However, Plan B for fighting global warming may be much more easily implementable, e.g., by ocean fertilization[25], than Plan A, carbon emission reduction, because Plan A requires unanimous coordination of many countries[5,26]. While the expected utility of Plan B may be higher, if we proceed with it we will have to abandon the Plan A, a message nobody wants to send in case of global warming as it discourages carbon restriction efforts. That is why discussions about geoengineering to mitigate global warming are not popular.

The main feature of such planning is that the one must prepare all lines of defense simultaneously, but implement them one by one, i.e., be able to go immediately to Plan B if it appears that Plan A is failing. I also should underscore that “Plan A, Plan B” is a useful concept that helps to structure different prevention ideas, but the reality of X-risks prevention is more complex.

## 2.2. Time stages of the realization of a plan

Any prevention plan will include similar stages of realization, like planning and funding. The main difference for global risks prevention plans is that we live in a constantly changing world, so plans that have a longer timescale may have to rely on the future advanced technology, but also deal with higher uncertainty. This is the list of the steps of each plan:

* Step 1 is developing an understanding of the nature of the risks and creating a prevention theory.
* Step 2 is preparation, which would include promoting the core idea of the plan, gathering funding, and building infrastructure needed for risks mitigation. Step 2 cannot be completed successfully without Step 1.
* Step 3 is the implementation of the preventive measures on a low technological level, at the current level of technologies. Such low-tech measures (e.g., technology bans, video surveillance) are realistic, but limited in scope. They could start to work sooner and cost less than high-tech measures from the next Step 4, so the chances of their implementation are higher as they start earlier and would provide protection for a longer period of time.
* Step 4 is the implementation of advanced measures based on the future technologies, which would finally end most risks, but which may themselves pose risks, like creation of an AI Nanny, or of a Nanoshield[27] to protect against nanobots.
* Step 5 is the final state, where our civilization would attain “indestructibility”[28], that is, an almost zero probability of global catastrophes, allowing our civilization to exist for an indefinitely long period.

Simpler versions of some projects could be created more quickly and provide some defense, but more complex versions would require more time and much more extensive funding. For example, when Turchin and Green analyzed the use of submarines as refuges in a global catastrophe, they counted eight stages, with costs ranging from one million to one hundred billion dollars, and timescales of preparation ranging from 5 to 30 years[18].

It is also clear that simpler versions of the projects are cheaper, so they have a higher probability of realization and shorter timing, giving a longer duration of protection and bigger chances of realization. They also have lower opportunity costs, as they attract less money, which otherwise could be spent on other projects. However, low-tech projects could be affected by future technological development, which could make these projects obsolete as new risks as well as new prevention technologies appear.

Some projects would not provide an opportunity for a second attempt. For example, a world-governing superintelligent AI could be made only once. If such an attempt were to fail, it would be catastrophic, because non-benevolent AI will prevent the appearance of any other AIs and may even cause human extinction.

# 3. Plans A and B for global risks prevention

## 3.1. Overview of global risk prevention plans

In the case of the X-risks, Plan A is prevention of the global risks and Plan B is survival of a possible global catastrophe. In Plan A the concern would be with the root of the problem, and in Plan B with its consequences. There could be some intermediate steps, like attempts to limit the size of a catastrophe or to slow the speed of its spread - not perfect solutions, but a potentially helpful component when survival may be at stake.

If we continue with that “causal” approach to global risk prevention plans, we could also identify a Plan C, which would be to leave information records about humanity after a catastrophe if nobody survives. Such traces could be used by the next civilization on Earth—if it ever appears—to revive humans; such a time-capsule could be placed on the Moon[29].

I also distinguish other two groups of plans. One group, “improbable plans,” are plans with an extremely low *a priori* probability of success, like sending messages to aliens asking for help, or that are based on low-probability speculative assumptions.

The other group consists of “bad plans,” plans with *a priori* negative utility, as they would result in a catastrophe rather than preventing it. Examples of the bad plans are depopulation, deliberate small catastrophes to stop progress, and Luddism[30,31].

So, based on the multilevel defense approach, there are five groups of plans for global risks prevention:

1. Plan A – prevent the catastrophe
2. Plan B – survive the catastrophe
3. Plan C – leave informational records about humanity
4. Plan D – improbable plans
5. Bad plans

I think that in the case of global risks, causal order is the same as expected utility order, being simpler and more probable to prevent X-risks than to merely survive them (more in Section 4).

It is important to stress that we might do well to prepare all plans simultaneously, e.g., building a Mars colony while working on the friendly AI on Earth. Plan B becomes active, if Plan A fails, but by preparing all plans simultaneously may help prevent multilevel defense against global risks.

## 3.2. Plan A: prevent the catastrophe

The biggest global risks are connected with the future development of the new technologies. Global risks may appear when a dangerous technology is obtained by a dangerous actor: a terrorist who obtains a dangerous virus, or a mad scientist, who starts a dangerous experiment, or a rogue dictator who starts to build a Doomsday machine, or an AI programmer who overestimates his ability to control AI. In the case of natural risks like a supervolcano eruption[32], I assume “zero actor” or a human inability to predict and prevent the risk.

Thus, there are two classes of “objects,” actors and technologies, and the goal of X-Risks prevention is to ensure that dangerous technologies will not get into dangerous hands[33], that is, to prevent the meeting of the objects from both classes.

Prevention becomes increasingly difficult with exponential technological progress, as technology becomes cheaper and more dangerous, and the number of actors grows. These advances make it possible for larger numbers of dangerous technologies to get into larger numbers of dangerous hands. Prevention requires stronger control over technologies and actors, and most ideas about how to prevent X-risks are centered about how to create the necessary level of monitoring.

For all possible potential sources of a catastrophe to be tightly observed, the level of control required to prevent it will need to be higher than the level of accidents. Here, control is defined as the ability to monitor, envision and act to address multiple potential and actual sources of risk.

1. Top-down approaches to global risk prevention mainly focus on how control (consisting of passive information gathering and ability to act) can be implemented to the necessary degree over the entire Earth. If controlling measures were not equally distributed globally, it would create “havens” for potentially dangerous activity. Nation-states differ in the level of control they exercise over their territory, and in some cases they themselves need to be controlled, for example, rogue states, failed states and “wannabe” superpowers.

The main question of control would be in how it can be globally implemented. What is the correct version of control? How can we distinguish that a given scientist is working on a safe viral vector versus a dangerous flu virus?

1. There are two types of global control solutions: peaceful, “evolutionary” solutions, and revolutionary solutions. Revolutionary solutions are *a priori* riskier, and they could be regarded as different variants inside Plan A.
2. **Peaceful solutions:**
3. 1. **International control**, something like a super United Nations, a collaboration of all nations to stop global risks. Such a collaboration should be established in several steps, starting with scientific research, then moving to international laws, then to a global risk prevention authority, and finally to “active shields” similar to high-tech immune system[27,34], using future super technologies to control other technologies.
4. The main problem with this approach would be in the difficulty of uniting the planet (without wars) and the uncontrollable nature of any world government that may appear as a result of the unification. Large human institutions often have low predictive ability and thus tend to overuse force to prevent risks. Greater insight[35] facilitates earlier action; thus, less force is needed to implement actions.
5. Many global risk prevention efforts are already operated by governments, like the Near-Earth Orbit (NEO) asteroid search program run by the US government[36], and the international cooperation during the recent Ebola epidemic[37], as well as the international anti-ISIS coalition. Thus, governmental collaboration is the most established method of global risks prevention. It is most suitable for controlling well-known risks, which are relatively easy to predict and manage.

In the future, such collaborations will include more and more elements of AI, like facial recognition in surveillance systems already happening in China[38], internet monitoring for signs of dangerous activity, and “world model” calculations, based on social networks data[39].

1. 2. **Citizen control or decentralized monitoring.** This is where the idea of openness comes into play. Distributed net systems have proved their effectiveness in many domains, such as blockchain, which records all Bitcoin transactions. The proposals for transparent society from Brin[10] and “OpenAI” from Musk[40] are members of this class of ideas. A decentralized monitoring approach includes possible changes of prevailing human values in a more risk-averse and earth-preserving way via advertising and education. Also, decentralized control is better for controlling new risks, as governments are too inert to adapt to them.

Non-self-improving (or slowly self-improving) AI will be helpful at all stages of such a project, as it will gather information, create future models, and control prevention systems. A network of powerful AIs would be even more helpful[41].

The main problem of decentralized monitoring is its weakness in enforcement over actors who choose to maintain secrecy and sovereignty.

**Revolutionary methods:**

1. 3. **Strong Friendly AI takeover.** If the problem of X-risks prevention is a problem of control, and control is mainly about correctly predicting the future, then we could say that X-risks prevention is an informational problem. In that case, powerful AI is necessary to manipulate information and create the best future models. (Here we have some circularity, as previously seen with world government: the best instrument of risk prevention is itself the main risk.) Yudkowsky and Bostrom hold the opinion that if we invest in the creation of friendly AI, it will help us to prevent all other risks[9,42]. Similarly, Ben Goertzel suggested creating an AI Nanny, which will be intrinsically safe and help us to manage other global risks[43].
2. We could create a non-self-improving and non-sentient AI-system[44] solely to help us fight other risks. As AI is the biggest X-risk and international cooperation is ineffective to fight it, this approach seems attractive. However, to take over the world, AI must be military AI on some stage[45]; it would also have to work from the first try, which is very unlikely. We should also remember some governments' attempts to create AI-like systems using global surveillance, like the NSA's PRISM, which harvests data from the internet[46], or private data analyzer “Palantir”[39].
3. For risk prevention in all scenarios in the foreseeable future, a global superintelligence control system provides the most stable long-term solution, as it will be able to predict all possible catastrophes far in advance.
4. 4. **Totalitarian world government appearing after a world war**, **or world takeover using a “superweapon.”** This idea is least appealing and many think that eternal totalitarianism is itself an existential risk[4]. Depending on its realization, it may belong in the “bad plans” section. This plan is risky for many reasons. World war could result in the use or development of new dangerous weapons; in addition, such a totalitarian government is not self-regulating and is subject to unlimited corruption and catastrophic mistakes. Such a government would probably use AI for “computer totalitarianism” and ubiquitous invasive control, including thought control and constant video surveillance. I want to underline that I am not advocating for it; I list it only to show that it is one possible, but undesirable, solution to the problem of the existential risks.

There are several auxiliary prevention solutions that do not require any global form of control, like differential technological development[4,47,48] or specific solutions for specific risks. It has been suggested that the growth of wisdom in society, the ability to envision future consequences of new technologies, should outperform technological development[47], and that such ability could be alternative to invasive totalitarian control.

All variants of Plan A result in almost the same end point: a global control system based on some form of AI and global surveillance. This system produces enough surveillance data and has enough predictive power to prevent all possible X-risks from appearing. The main difference is whether the system would be based around a single actor, i.e., superintelligence, around large arrays of sensors with narrow AI in the center, or a distributed net of AIs.

## 3.3. Plan B: survive the catastrophe

Plan B is to survive a catastrophe resulting from an X-risk. There are several means of survival, but they fall into three main categories: improving resilience, escape, or hiding.

**Improving resilience** means lessening the scale of the potential catastrophe or increasing survivability of structures and human beings[49–51]. It also includes a larger population, more diversity of habitats and lifestyles, better medicine and stockpiling of goods.

**Escape** in this context can only mean space colonization. The first step is temporary asylums in space, like the International Space Station (ISS)[15], and the second is colonizing the Moon[52] and Mars[8]. Higher level technologies based on robotics and replicators will help us to eventually colonize asteroids, and the final step will be interstellar humanity. Unfortunately, space colonization comes with the new risk of space wars[53].

**Hiding** would require creation of temporary asylums on Earth, like underground bunkers, ships, and Antarctic settlements. They could help only in a small range of catastrophe types, like medium-size asteroid impacts or some bio risks. The best possible solution here is the use of old nuclear submarines[18]. Islands and other types of refuges have also been discussed[19].

While implementation of the Plan B starts only after Plan A’s (possible) failure, preparation for all plans must take place simultaneously. Such preparation could be performed by various agents independently in the earlier stages of the projects but would require collaboration in later stages.

## 3.4. Map of X-risk prevention plans

Table 1 illustrates the most simplified version of all currently known X-risks prevention plans. (The full version of the map can be found online at http://immortality-roadmap.com/globriskeng.pdf)

One of the goals of the map is to collect all possible ideas in one place; to that end, around 100 ideas were evaluated. Many ideas for this map were crowdsourced via the LessWrong forum and other social networks in 2015, and most of the ideas suggested by the public are presented in the block “decentralized risks monitoring”. A total of 21 ideas were added to the map based on the crowdsourcing.

*Table 1. Overview of X-risks prevention plans (see full version here:* [*http://immortality-roadmap.com/globriskeng.pdf)*](http://immortality-roadmap.com/globriskeng.pdf%29) *.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | Subplans | Stage 1Research | Stage 2 Pre-paration | Stage 3Low-level implement-ation | Stage 4 High-level implementation |
| Plan A.Prevent catastrophe | **Plan A1** Peaceful integration | ***International cooperation*** | Risk modelling | Social support | Super UN | High tech shields |
|  |  | ***Horizontal monitoring*** | Value changes | Improving humans | Social changes | Decentralized control |
|  | **Plan A2**. World takeover | ***Beneficial super-intelligence***  | Promotion of AI safety | AI alignment theory | Slow down creation of strong AI by other actors  | Strong AI becomes global superintelligence |
|  |  | ***World government*** |  |  | Planet unification war | Global surveillance system |
|  | **Plan A3**.  | ***Lowering risks*** |  | Quaran-tine | Specific solutions for concrete risks | Technological speed-up |
| Plan B. Survive | **Plan B1**.  | ***Resilience*** | Food stockpiles | Improving sustainability | Raise resilience of human bodies | Miniaturization and virtualization |
|  | **Plan B2.**  | **Space colonization** | Space station | Moon and Mars | Solar system | Interstellar travel |
|  | **Plan B3.**  | ***Refuges*** | Bunkers | Military submarines | Ark-submarines | Nanotech-based refuges |
| Plan C | **Leave information about humanity** |  | M-disks | Time-capsule on Moon | Space memorials | Nanobots in space |
| Plan D | **Improbable ideas** |  | Messages to aliens |  |  |  |
| Bad ideas |  |  | Depopulation | Controll-able regression | Hope or ignore | Stop X-risks research |

# 4. Navigating prevention plans’ space

## 4.1. Utility and probability of success for different plans

I classified X-risks prevention plans based on their position in the causal chain of events, resulting in a large spectrum of possible plans in the form of a map (Table 1). The next step is to choose a path within this map.

There are three main questions to facilitate this decision:

1. Which version of Plan A is most effective? We have four main approaches and several auxiliary approaches; they will contradict each other at later stage of their realization.
2. Could Plan B be more effective than Plan A after we calculate all costs and probabilities?
3. Which combination of plans is most effective?

To answer these questions, I needed a way to estimate the expected utility of each plan, that is, measure its usefulness. The expected utility of an X-risks prevention plan should be measured through the total probability change of global catastrophic risks on dollar spent. I should note the intrinsic difficulty of applying the notion of probability to such unprecedented one-time events as global catastrophes[54].

This expected utility estimation is based on several elements:

* **Probability of success** if the plan is implemented.
* **Opportunity cost**, measured as a probability that other plans will not work because of realization of this plan. For example, technological relinquishment would prevent space colonization.
* **Collateral global risks** connected with the plan, that is, the probability that the prevention plan itself becomes an X-risk. For example, a Nanoshield could go awry, world government could become a totalitarian dictatorship with stupid and risky behavior, friendly AI takeoff could be the beginning of AI wars, space stations could give rise to space weapons, etc.

There are other variables that typically affect the expected utility of any action; in our case, they will affect the chances that decision makers will take the given course of action, so they could be converted into a combined probability of plan implementation. A recent article by Maas et al. explores other aspects of X-risks, such as vulnerability and exposure[55]. In a related work, Matheny has explored the cost-effectiveness of asteroid risks prevention[56]. These other variables, which will affect decisions of the policymakers are:

* **Moral cost**, as the number of lives lost during the plan’s implementation.
* **Time** required for the plan’s realization.
* **Price** in USD or as a percent of GDP.
* **Additional benefits**, like prevention of smaller catastrophes.
* **Proof**, as the level of proof that this is exactly the plan needed and it is an effective preventative measure for real X-risks.

In order to achieve maximum global risks protection, we need to choose a combination of several plans from the categories of Plan A and Plan B, so they do not contradict each other and most effectively make use of available resources.

If we calculate the total probability of success, it may not be as high for Plan A as one might expect because of uncertainty and possible collateral damage. Some prevention plans are difficult to implement because while they may have high chances of success, like international government and friendly AI, they may also have high collateral risks, as both could fail catastrophically. As a result, their expected utility may not be as high as one may imagine.

Some Plan B variants may be cost-effective and not incur any additional risks. They may be easy to implement and have fewer risks associated with them, but their chances of success are low. For example, an attempt to use nuclear submarines as survival refuges is effective and doesn’t pose any additional risks. As a result, Plan B may have a higher chance of working than Plan A in certain situations, and have higher cost-effectiveness. However, given limited total funding for X-risks prevention and limited amount of public attention, more careful choices are needed, which will not exhaust these precious recourses.

I could estimate that the probability of success of some version of Plan A is 10 percent, Plan B – 1 percent, and Plan C – below 0.01 percent. This is my estimation of the probability that the realization of this plan will prevent a global catastrophe on the condition that no other plan has been implemented, and that the catastrophe is inevitable if no prevention plans exist at all. It is important to consider that bad plans will increase the likelihood of a global catastrophe rather than reducing risk.

## 4.2. Individual plan’s efficiency

Although each plan is suitable for preventing all possible risks, each particular plan is also most efficient for preventing of a certain type of disaster. For example, Plan A1.1 (international control system) is best suited to controlling the spread of nuclear, chemical, biological weapons and anti-asteroid protection, whereas Plan A2.1 (beneficial AI superintelligence) is best for preventing the creation of dangerous AI, as well as nanotech risks and advanced biotech risks, for they are less anticipated by existing governmental systems and require more complex prevention measures.

In another example, space exploration is well-suited to protect against asteroids but does very little to protect against a dangerous AI that can travel via communication lines. This underlines the importance of multilayered defense and the integration of multiple plans. It also shows that different plans are better on different time horizons and for different models of the future. An international control system is better for shorter time horizons or for low-tech, stagnating future, while AI is better for controlling an exponentially exploding future.

## 4.3. Strategies to realize prevention plans

After creating the X-risks prevention map and evaluating the expected utility of each approach, we can evaluate possible paths to global safety as to which would be the most effective combination of plans. There are three main approaches to such a path:

1. The **all-in approach**, when one plan seems to have a higher expected utility than all others plans combined. For example, Yudkowsky thinks that beneficial superintelligent AI[9] is the possible solution to all other problems.

2. **Multilevel defense approach**, in which humanity works on all plans simultaneously. This approach is useful if many plans have similar expected utility, but are independent and could be created by independent actors, each with their own funding.

3. **Some synergic combination of plans**, where elements of one plan help the realization of another, as in Musk’s strategy (AI research + use of solar energy + establishing a Mars colony)[8,57,58].

Individual actors may choose individual plans for which they go all-in. If there isn’t a global authority (e.g., world government or superintelligent AI) to choose the right plan, surely the result would be a combination of plans, presented by different competing actors.

Most of the plans suggested above are not mutually exclusive in their early stages. The main factor that may make them exclude each other is the question of the global power over the Earth: a super UN, AI, a union of strong nations, team of hackers, one country, or a decentralized civil risk monitoring system. This issue will only arise at later stages of plan realization.

This question of “global power” is so serious that it is itself a major global risk, as there are many entities eager to take power over the world. These entities could start a world war by fighting against each other. Basically, this becomes a question of choosing which variant of Plan A to implement. There is no simple answer, but I hope that peaceful integration would dominate.

## 4.4. Musk’s Plan A and Plan B

As the most complex of publically known and actor-supported plans of global risks prevention are Musk’s set of existential risks prevention ideas, they require deeper analysis.

Musk’s Plan A is solving AI safety via OpenAI[40] and human augmentation via brain-computer interface(BCI), “Neuralink”[7], while in parallel solving the problem of global warming by introducing electric cars and solar energy. His Plan B is a colony on Mars[57] with one million people. Musk also put a disk with data about humanity from Arch Mission on his Falcon Heavy first launch, which could be regarded as an element of Plan C, that is preservation data about humanity[59]. These plans help each other by providing mutual funding and sharing technologies.

However, there were some doubts expressed about whether this exact combination of plans would be the most optimal approach. Bostrom wrote about his misgivings that openness in AI development and human augmentation would be the best approaches for AI safety[60]. Like fossil fuels, solar energy also has environmental costs[61]. A space colony on Mars would be very expensive, fragile, and provide protection from a very small range of risks, and could be subject to international conflicts[62], and disks with data would be more secure on the Moon, with a higher chance that they may be found there by another civilization.

## 4.5. Combining plans in the most safe and cost-effective way

From my point of view, the best Plan A would start by raising awareness about global risks, with the hope of facilitating effective international cooperation coupled with decentralized monitoring. This may be followed by the creation of a strong “super UN” based on world democracy, with a mandate to fight global problems. In later stages, it should be empowered by a non-self-improving non-superintelligent AI Nanny[44], a narrow AI built for surveillance and prediction of global risks. This plan would not exclude Musk’s private project; in fact, the plans could be coupled.

The optimal Plan B is the use of nuclear submarines[18], remote islands[19] and terrestrial bunkers as low-tech and rather cheap refuges from a variety of global risks, with later use of self-replicating robots to explore and terraform our solar system for a relatively low price.

The most effective Plan C is to put data under the surface of the Moon in pole regions with information about humanity on engraved metal disks and include storage for frozen DNA samples. This would be combined with beacons on the surface of the Moon designed to survive for tens of millions of years to be easily found by the next civilization to arise on Earth, or even by aliens[29].

# 5. Conclusion

As of 2018, the X-risks prevention situation is dire, as there are more than 10 global risks with different probabilities. Furthermore, there is not much consensus about them as they are covered by the fog of multilevel uncertainty. The prevention strategy is in even worse condition:

* There is no globally accepted universal picture of X-risks.
* No prevention plan is accepted by the international community.
* Each possible prevention plan carries its own risks.
* There is no consensus of how these plans could be implemented, from both organizational and funding points of view.
* The X-risks community has many ideas, but we don’t know which are the best, nor do we know how to go from the current situation to a practical realization of any of the mentioned plans.
* The suggested plan has no agent. It describes what humanity as a whole could do for X-risks prevention, but humanity is not currently united. Individuals or state actors have to find their place in this plan.

Despite these limitations, to have a plan (providing some orientation in the space of possibilities) is better than to have no plan at all. Like any framework, the suggested two-dimensional map of plans would have its limitations, as some of the ideas may not properly fit into it, or may be difficult to classify.

**Alexey Turchin is the author of several books and numerous articles on the topics of existential risks and life extension, as well as a series of roadmaps on the same topics. He is a vice president of the Foundation Science for Life Extension and the founder of the startup Digital Immortality Now. His articles can be seen at** [**https://philpapers.org/profile/654846**](https://philpapers.org/profile/654846)**, and his email is** **alexeiturchin@gmail.com**

### References

[1] R. Kurzweil, Singularity is Near, Viking, 2006.

[2] R. Freitas, Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations, Foresight Institute Technical Report, 2000.

[3] P. Millett, A. Snyder-Beattie, Existential Risk and Cost-Effective Biosecurity, Health Secur. 15 (2017) 373–383. doi:10.1089/hs.2017.0028.

[4] N. Bostrom, Existential risks: Analyzing Human Extinction Scenarios and Related Hazards, J. Evol. Technol. Vol 9 No 1 2002. (2002).

[5] N. Bostrom, Existential risk prevention as global priority, Glob. Policy. 4 (2013) 15–31.

[6] M. Rees, Our Final Century, Heinemann, 2003.

[7] T. Urban, Neuralink and the Brain’s Magical Future, (2017). http://waitbutwhy.com/2017/04/neuralink.html.

[8] H. Devlin, Life on Mars: Elon Musk reveals details of his colonisation vision, The Guardian. (2017). http://www.theguardian.com/science/2017/jun/16/life-on-mars-elon-musk-reveals-details-of-his-colonisation-vision (accessed April 4, 2018).

[9] E. Yudkowsky, Artificial Intelligence as a Positive and Negative Factor in Global Risk, in Global Catastrophic Risks, Oxford University Press: Oxford, UK, 2008.

[10] D. Brin, The Transparent Society, Perseus Book, 1998.

[11] R. Posner, Catastrophe: Risk and Response, 2004.

[12] D. Denkenberger, J.M. Pearce, Feeding everyone no matter what: managing food security after global catastrophe, Academic Press, 2014.

[13] K. Jebari, Existential risks: Exploring a Robust Risk Reduction., Sci. Eng. Ethics 21 541 Doi101007s11948-014-9559-3. (2014).

[14] T.M. Maher, S.D. Baum, Adaptation to and recovery from global catastrophe, Sustainability. 5 (2013) 1461–1479. doi:doi:10.3390/su5041461.

[15] P. Torres, Space bunkers, (2016).

[16] N. Beckstead, How Much Could Refuges Help Us Recover from a Global Catastrophe?, Futures. 72 (2015) 36–44.

[17] A. Turchin, The Map of Shelters and Refuges from Global Risks, Plan B of X-risks Prevention, 2016. altruism.com/ea/12x/the\_map\_of\_shelters\_and\_refuges\_from\_global\_risks/.

[18] A. Turchin, B.P. Green, Aquatic Refuges for Surviving a Global Catastrophe, (2017).

[19] A. Turchin, B.P. Green, Islands as refuges for surviving global catastrophes, Submited GoCAS Foresight Spec. Issue. (2018).

[20] C. Goldblatt, A.J. Watson, The runaway greenhouse: implications for future climate change, geoengineering and planetary atmospheres, Phil Trans R Soc A. 370 (2012) 4197–4216. doi:10.1098/rsta.2012.0004.

[21] W.E. Burrows, The survival imperative: Using space to protect earth, Collect. Tech. Pap. - 2004 Planet. Def. Conf. Prot. Earth Asteroids. (2004) 548–563.

[22] G. Petrangeli, Nuclear Safety, Elsevier, 2006.

[23] H. Kahn, On thermonuclear war, Princeton University Press, 1959.

[24] M. Inman, Planning for plan B, (2010). http://www.nature.com/climate/2010/1001/full/climate.2010.135.html.

[25] A.J. Watson, C.S. Law, K.A.V. Scoy, F.J. Millero, M.I. Liddicoat, R.H. Wanninkhof, R.T. Barber, K.H. Coale, Minimal effect of iron fertilization on sea-surface carbon dioxide concentrations, Nat. 371 (1994) 143–145.

[26] N. Bostrom, The Unilateralist’s Curse: The Case for a Principle of Conformity, (2012). http://www.nickbostrom.com/papers/unilateralist.pdf.

[27] V. Freitas, Nanoshield project, in: 2006. https://lifeboat.com/ex/nano.shield.

[28] A. Kononov, The goal of the undestructability of the civilization in the catastropchically unstable world., Pbliaction Inst. Syst. Anal. T.31 (2007).

[29] A. Turchin, D. Denkenberger, Surviving global risks through the preservation of humanity’s data on the Moon, Acta Astronaut. (2018) Pages 161-170.

[30] Kaczynski, Industrial Society and Its Future, The New York Times, 1995.

[31] R. Hanson, Overcoming Bias : Kaczynski’s Collapse Theory, (2018). http://www.overcomingbias.com/2018/01/kaczynskis-collapse-theory.html (accessed February 13, 2018).

[32] D.C. Denkenberger, R.W. Blair, Interventions that May Prevent or Mollify Supervolcanic Eruptions, Futures. (2018).

[33] P. Torres, Agential Risks: A Comprehensive Introduction. 2016, J. Evol. Technol. -. 26 (2016).

[34] S. Maurer, Bioshield project, (2006). http://lifeboat.com/ex/bio.shield.

[35] L. Muehlhauser, A. Salamon, Intelligence explosion: Evidence and import, in: Singul. Hypotheses, Springer, 2012: pp. 15–42.

[36] NASA, Center for NEO Studies, (2018). https://cneos.jpl.nasa.gov/ (accessed May 7, 2018).

[37] T.K. Mackey, The Ebola Outbreak: Catalyzing a “Shift” in Global Health Governance? BMC Infect. Dis. 16 (2016). doi:10.1186/s12879-016-2016-y.

[38] S. Liao, Chinese police are expanding facial recognition sunglasses program, The Verge. (2018). https://www.theverge.com/2018/3/12/17110636/china-police-facial-recognition-sunglasses-surveillance (accessed May 7, 2018).

[39] A. Winston, Palantir has secretly been using New Orleans to test its predictive policing technology, The Verge. (2018). https://www.theverge.com/2018/2/27/17054740/palantir-predictive-policing-tool-new-orleans-nopd (accessed April 10, 2018).

[40] G. Brockman, I. Sutskever, Introducing OpenAI, (2015). https://openai.com/blog/introducing-openai/.

[41] A. Turchin, D. Denkenberger, Global Solutions of the AI Safety Problem, (2017).

[42] N. Bostrom, Superintelligence, Oxford University Press, Oxford, 2014.

[43] B. Goertzel, Should Humanity Build a Global AI Nanny to Delay the Singularity Until It’s Better Understood?, J. Conscious. Stud. 19 No 1–2 2012 Pp 96–111. (2012). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.352.3966&rep=rep1&type=pdf.

[44] A. Turchin, Narrow AI Nanny: Deceive Strategic Advantage via Narrow AI to Prevent Creation of the Superintelligence, Manuscript. (2018).

[45] A. Turchin, D. Denkenberger, Military AI as convergent goal of the self-improving AI, Artif. Intell. Saf. Secur. Roman Yampolskiy Ed CRC Press. (2018).

[46] B. Gellman, L. Poitras, U.S., British intelligence mining data from nine U.S. Internet companies in broad secret program, Wash. Post. (2013). https://www.washingtonpost.com/investigations/us-intelligence-mining-data-from-nine-us-internet-companies-in-broad-secret-program/2013/06/06/3a0c0da8-cebf-11e2-8845-d970ccb04497\_story.html (accessed May 8, 2018).

[47] L. Muehlhauser, Facing the intelligence explosion, Mach. Intell. Res. Inst. Berkley Calif. (2013).

[48] B. Tomasik, Differential Intellectual Progress as a Positive-Sum Project, (2013).

[49] D.P. Aldrich, S. Oum, Y. Sawada, Approaches towards effective disaster risk-Coping strategies and regional cooperation on disaster management, Resil. Recovery Asian Disasters Community Ties Mark. Mech. Gov. (2015) 339–353. doi:10.1007/9784431550228\_16.

[50] S.D. Baum, Resilience to global food supply catastrophes., in: Environ. Syst. Decis. 352, 2015: pp. 301–313.

[51] S.D. Baum, Risk and resilience for unknown, unquantifiable, systemic, and unlikely/catastrophic threats, Environ. Syst. Decis. 35 (2015) 229–236. doi:10.1007/s10669-015-9551-8.

[52] J. Pass, Moon bases as initial “space society” trials: Utilizing astrosociology to make space settlements livable, AIP Conf. Proc. 880 (2007) 806–813. doi:10.1063/1.2437520.

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

[54] A. Turchin, D. Denkenberger, Global Catastrophic and Existential Risks Scale, Futur. Press. (2018).

[55] H.-Y. Liu, K.C. Lauta, M.M. Maas, Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research, Futures. (2018).

[56] J.G. Matheny, Reducing the risk of human extinction, Risk Anal. 27 (2007) 1335–1344. doi:doi:10.

[57] E. Musk, Making Humans a Multi-Planetary Species, New Space. 5 (2017) 46–61.

[58] A. Chen, Elon Musk’s dreams of merging AI and brains are likely to remain just that — for at least a decade, The Verge. (2017). https://www.theverge.com/2017/4/21/15370376/elon-musk-neuralink-brain-computer-ai-implant-neuroscience (accessed February 14, 2018).

[59] Arch Mission, FAQ, Arch Mission. (2018). https://www.archmission.com/faq/ (accessed February 11, 2018).

[60] N. Bostrom, Strategic Implications of Openness in AI Development, Work. Draft. (2016). http://www.nickbostrom.com/papers/openness.pdf.

[61] Union of Concerned Scientists, Environmental Impacts of Solar Power, Union Concerned Sci. (n.d.). https://www.ucsusa.org/clean\_energy/our-energy-choices/renewable-energy/environmental-impacts-solar-power.html (accessed May 8, 2018).

[62] K. Szocik, T. Wójtowicz, L. Baran, War or peace? The possible scenarios of colonising Mars, Space Policy. (2017).