The Great Downside Dilemma For Risky Emerging Technologies

Seth D. Baum, Global Catastrophic Risk Institute
http://sethbaum.com * http://gcrinstitute.org

Physica Scripta, 89(12):128004 (2014). This version dated 27 November 2014.

Abstract
Some emerging technologies promise to significantly improve the human condition, but come with a risk of failure so catastrophic that human civilization may not survive. This article discusses the great downside dilemma posed by the decision of whether or not to use these technologies. The dilemma is: use the technology, and risk the downside of catastrophic failure, or do not use the technology, and suffer through life without it. Historical precedents include the first nuclear weapon test and messaging to extraterrestrial intelligence. Contemporary examples include stratospheric geoengineering, a technology under development in response to global warming, and artificial general intelligence, a technology that could even take over the world. How the dilemma should be resolved depends on the details of each technology's downside risk and on what the human condition would otherwise be. Meanwhile, other technologies do not pose this dilemma, including sustainable design technologies, nuclear fusion power, and space colonization. Decisions on all of these technologies should be made with the long-term interests of human civilization in mind. This paper is part of a series of papers based on presentations at the event Emerging Technologies and the Future of Humanity held at the Royal Swedish Academy of Sciences, 17 March 2014.

Keywords: risk, emerging technologies, nuclear weapons, extraterrestrial intelligence, geoengineering, artificial intelligence

1. Introduction

Would you play a game of Russian roulette? Would you take a six chamber revolver, put one bullet in, give it a spin, point it at your head, and pull the trigger? How about for a million dollars? Would you play?

I would guess that most readers of this paper would not play. I would guess that you would think that a chance of one million dollars is not worth it to take this risk of ending up with a bullet in the brain. I personally would not play, for the same reason. Our brains and our lives are simply worth more than that.

But suppose your life circumstances were different. Suppose you were struggling with money, that you were basically broke. Suppose you were sick, with a chronic condition you cannot afford to cure. Maybe you do not have that many years left to live anyway. Now, with less to lose and more to gain, that game of Russian roulette might start to look more attractive. Now, you might start counting how much that million dollars could do for you. Could it cure your sickness, make you healthy again? Could it add years to your life? Could it pull you out of poverty? Could it give you basic comfort?

If the million dollars would do enough for you, then maybe you would choose to play. Desperate circumstances can sometimes warrant taking desperate risks. If it works, the circumstances get better, maybe much better. But it might not work, and if it does not, it comes

with a downside—in this case, a bullet in the brain. Whether to play the game is a downside dilemma: a dilemma involving a significant possible downside.

This paper talks of a *great* downside dilemma. It is great because the stakes are so high—indeed, they are literally astronomical. At stake is not the fate of a single person, as in Russian roulette, but the fate of human civilization. This includes the roughly seven billion people alive today and the many more members of all the future generations that might ever live. The stakes are astronomical because humans (or our descendants) might be able to colonize space and achieve great things across the universe. Human civilization already has an active space program, and space colonization seems feasible, as long as no great catastrophe denies humanity the chance. The rest of the universe is vastly larger than our humble home planet, so space colonization would open up enormous new opportunities. Meanwhile, for all humanity currently knows, humans might have the only intelligent civilization anywhere in the universe. And so the stakes could mean nothing less than the success or failure of intelligent civilization in the entire universe. A *great* downside dilemma, indeed.

To be more specific, the great downside dilemma is any circumstance in which human civilization must choose whether to take a risk in which, if it works out, the benefit greatly improves the human condition, but if it does not work out, a catastrophe will occur, a catastrophe so large that civilization could perish, a metaphorical bullet in the brain. The dilemma is whether to take the risk. How much does civilization value that improvement in its condition? Could it be enough to pull civilization out of desperate circumstances? How large is the risk of catastrophe? Is it small enough that the risk is worth taking? Can any risk of civilization perishing be small enough to justify taking the risk? These questions must be answered in order to decide whether to take the risk.

The great downside dilemma arises often for decisions about whether to pursue certain emerging technologies. These technologies promise to solve major societal problems. They bring peace, cure disease, protect the environment, and more. Or rather, they do these things if they work as intended. However, they may not work out as intended. They may fail, and fail catastrophically. In the worst cases, they could kill every living human—the extinction of our species—and destroy much of the rest of Earth's biosphere as well. Should society develop and launch these technologies, given their promise and despite their risks? That is the great downside dilemma for emerging technologies. This dilemma is an important issue for society as a whole and especially for scientists and engineers, who by virtue of their background are especially able to contribute to the debate.

This dilemma is one important part of the broader challenge of avoiding civilization-ending global catastrophes. A growing body of scholarship recognizes the avoidance of these catastrophes as crucial for the long-term success of human civilization, and likewise as a key priority for action today [1-8]. Visionary technologist James Martin likened this era of civilization to a turbulent river that it must navigate [9]. If this era of civilization successfully navigates the river, then a long, bright future awaits, both on Earth and beyond. However, if it fails, then human civilization suffers a premature death. This paper describes several great downside dilemmas for emerging technologies and explains how humanity can navigate through them. The paper also discusses some other technologies that do not pose this dilemma because they promise to bring major benefits without a significant catastrophic risk.

2. Historical Precedents

Amazingly, the great downside dilemma for emerging technologies has been faced at least twice before. The first precedent came in the desperate circumstances of World War II. The dilemma was whether to test-detonate the first nuclear weapon. While nuclear weapons proved to be unprecedentedly destructive weapons, a single detonation did not destroy the entire planet as some initially feared. The second precedent came during calm circumstances but still posed a dilemma every bit as large: whether to engage in messaging to extraterrestrial intelligence (METI). METI is of note because the dilemma still has not been resolved. Humanity still does not know if METI is safe. Thus METI decisions today face the same basic dilemma as the initial decisions in decades past.

2.1 Nuclear Weapons

It was 1945, towards the end of World War II. An American team of physicists, engineers, and military personnel built the first atomic bomb, which they named Trinity. Trinity was to be detonated in a test explosion, to make sure the technology worked, before using additional atomic bombs against Japan. By that point, Germany had already surrendered. Japan was nearing defeat, and the United States believed that the atomic bomb could compel Japan to surrender without the U.S. waging a long, bloody invasion. It might seem counterintuitive, but this most powerful of weapons was built to save lives.[1]

However, some of the physicists worried that the test might fail catastrophically. They worried that the detonation could ignite the atmosphere, ending life on Earth. They believed the chance of this happening to be exceptionally small, due to their understanding of the relevant physics. Still, they closed their report on the topic with the line "However, the complexity of the argument and the absence of satisfactory experimental foundations makes further work on the subject highly desirable" [11]. Thus the risk did give them some pause. Sure enough, they took the risk. As is now known, the Trinity test succeeded: the bomb worked, and the atmosphere did not ignite. The rest is history.

Humanity survived the first atomic bomb detonation, the next two, which were dropped on Hiroshima and Nagasaki, and the 2,054 atomic bombs that have been detonated since in further testing.[2] The largest of these bombs, the Soviet Tsar Bomba, had a yield equivalent to about 50 megatons of TNT, a whopping 2,500 times larger than Trinity. The atmosphere did not ignite. And physics has by now progressed to the point where we understand with very high confidence why atomic bomb detonations would not cause these harms (though they can of course cause other harms). But for that brief moment in time, when the first atomic detonation was under consideration, a great downside dilemma was faced, without the benefit of hindsight that now exists.

Today, nuclear weapons remain a major threat. A single nuclear weapon could kill thousands or even millions of people. Nuclear war with hundreds or thousands of weapons (about 17,000 weapons still exist, mainly held by the United States and Russia) would not produce enough radiation to cause human extinction, as Bertrand Russell, Albert Einstein, and others once feared [12]. But it would cause significant cooling, as the smoke from burning cities rises into the atmosphere and blocks incoming sunlight. This cooling, often known as nuclear winter, could cause widespread agricultural failure, with the resulting famine threatening millions or even

[1] Reed [10] reviews the history of the atomic bomb development and the corresponding physics.

[2] Atomic bomb detonation data can be obtained from the Comprehensive Nuclear-Test-Ban Treaty Organization at http://www.ctbto.org/nuclear-testing/history-of-nuclear-testing/world-overview/page-1-world-overview/.

billions of people [13-14]. The worst case scenario could include human extinction [15]. The leaders of nuclear weapons states thus face a different dilemma: In a crisis, are nuclear weapons worth using? While nuclear weapons are no longer a new technology, large nuclear arsenals have never been used in war, and so the dilemma must still be resolved without the benefit of hindsight.

2.2 Messaging to Extraterrestrials

It was 1974, a relatively calm and ordinary year by most measures. But an unusual exercise was in preparation at an astronomy observatory in Arecibo, Puerto Rico. The Arecibo Observatory hosted what was (and still is) the largest radio telescope in the world.[3] Usually, the telescope is used either for radio astronomy, which detects radio waves incoming from the rest of the universe, or radar astronomy, which studies the solar system by sending radio waves towards planets and other nearby objects and analyzing the waves that bounce back. Radio and radar astronomy are generally harmless and of scientific value. But in 1974, the Arecibo telescope was to be used differently.

The plan was to send a message from Arecibo to a cluster of stars 25,000 light years away. The Arecibo Message was designed by astronomer Frank Drake and colleagues with the premise of METI. The message contained seven parts describing human physiology and astronomy. This was not the first exercise in METI. In 1962, the Morse Message was sent from the Evpatoria Planetary Radar in Crimea to Venus. But the Morse Message was as harmless as regular radar astronomy studying Venus. Because the Arecibo message was broadcast elsewhere, it broke new ground.

So far, the Arecibo message has not received a response. Of course it has not: it was sent 40 years ago to a location 25,000 light years away. It will take at least another 49,960 years for the message to arrive and the response to reach back to Earth.[4] And it is possible that no ETI will receive the Arecibo message. Indeed, it is possible that there are no ETI out there to receive it. It is also possible that ETI will receive it but not respond in any way. So even after another 49,960 years, the Arecibo message could prove inconsequential to humanity, except for its modest educational value. But it might not. The message could receive a response.

Despite some proclamations to the contrary, humanity has little understanding of how an encounter with ETI would proceed. Some people expect that ETI would benefit humanity, providing scientific knowledge and intercultural exchange, or even solutions to humanity's problems. Others expect that ETI would harm humanity, enslaving or even killing us. These are among the many possible outcomes of ETI encounter [16-17]. Presumably, if it was known that the outcome would be beneficial, METI would proceed; likewise it would not proceed if the outcome was known to be harmful [18]. But in 1974, it was not known. Whether to engage in METI thus posed a great downside dilemma.

In 2014, at the time of this writing, it is still not known whether METI is safe. Since the Arecibo message, several other messages to ETI have since been sent to closer stars, most recently the 2013 Lone Signal project.[5] None of these messages has yet received a reply.

---

[3] Arecibo is the world's largest single-aperture radio telescope. Arrays of multiple telescopes combined as astronomical interferometers collectively cover larger areas.

[4] Assuming the location is exactly 25,000 light years away, then 49,960 years from now is the minimum time it could take for a response to reach Earth. The response could reach Earth later if the ETI take more time before transmitting the response.

[5] Full disclosure: I received funds from Lone Signal to contribute to a risk analysis of Lone Signal's transmissions. The study concluded that the transmissions posed no significant risk because the transmitter Lone Signal was using

Meanwhile, it is an exciting time for SETI—the search for ETI. Astronomers are just starting to discover extrasolar planets [19]. But no ETI have yet been found. Until then, humanity will have deep uncertainty about the merits of METI. Some progress can be made by carefully thinking through the possibilities of ETI encounter. An argument can be made that no high-power METI should be conducted until humanity better understands the risks [20], but this is a controversial point. The great downside dilemma for METI persists.

3. Dilemmas in the Making

While humanity continues to face dilemmas related to nuclear weapons and METI, new dilemmas lurk on the horizon. The stakes for these new dilemmas are even higher, because they come with much higher probabilities of catastrophe.[6] I will focus on two: stratospheric geoengineering and artificial general intelligence. Neither technology currently exists, but both are subjects of active research and development. Understanding these technologies and the dilemmas they pose is already important and will only get more important as the technologies progress.

3.1. Stratospheric Geoengineering

In summer 2010, heavy monsoons flooded about one fifth of Pakistan, with millions of people affected [21]. The floods were part of a broader northern hemisphere summer heat wave that set temperature records in many locations. Vast wildfires in western Russia produced so much smoke that people in Moscow wore masks and airports redirected traffic. The floods, heat wave, and wildfires are among the sorts of extreme weather events that are expected to happen more often and with greater intensity as global warming worsens [22].[7]

The standard means of lessening the harms of global warming is to reduce atmospheric emissions of carbon dioxide, methane, and other greenhouse gases. That means using energy more efficiently, switching away from coal power, reversing deforestation, and more. But despite the risks of global warming, greenhouse gas emissions have been steadily increasing, and are projected to continue increasing into the future. More emissions means warmer temperatures and more severe harms from extreme weather events, sea level rise, and more.

In despair over the perceived failure to reduce emissions, observers are increasingly considering geoengineering to lower global temperatures. Geoengineering is the intentional manipulation of the global Earth system [27]. Greenhouse gas emissions do not qualify as geoengineering because they are an unintended byproduct of activities with other aims. Perhaps the most commonly discussed type of geoengineering is stratospheric geoengineering, which would lower temperatures by injecting particles into the stratosphere, thereby blocking a portion of incoming sunlight. Stratospheric geoengineering is attractive because of its relative feasibility, efficacy, and affordability. However, stratospheric geoengineering changes regional temperature and precipitation patterns, leaving some need to adapt to climatic changes. Stratospheric geoengineering also does nothing to address the acidification of oceans caused by carbon dioxide

---

at the time did not exceed the background radio wave leakage from radio and television broadcasts [18].

[6] This assumes that the probability of catastrophe from METI is relatively low. This is a debatable point, given the deep uncertainty surrounding the possibility of extraterrestrial contact.

[7] I am using the term "global warming" here instead of the usual "climate change" to distinguish it from nuclear winter, which is also a climatic change. Any readers who still doubt the legitimacy of global warming as an issue should consult some of the many works on the topic, including accessible books by leading global warming researchers [23-24] and my own humble contribution [25]. Global warming research is also voluminously reviewed by the Intergovernmental Panel on Climate Change [26].

emissions being absorbed into oceans. Ocean acidification is a major problem in its own right, a large threat to ocean ecosystems. Finally, stratospheric geoengineering also poses significant risks that could even exceed those of global warming.

Perhaps the largest risk from stratospheric geoengineering is the possibility of abrupt halt. Particles put into the stratosphere will cycle out on time-scales of about five or ten years. In order to maintain stable temperatures, particles must be continuously injected into the stratosphere. If the geoengineering abruptly halts, such that additional particles are not injected, then temperatures will rapidly shoot back up towards where they would have been without geoengineering [28]. This rapid temperature increase could be especially difficult to adapt to and thus be especially disruptive. For example, it may be difficult to determine which crops to plant in a given region, because the crops suitable for that region will change too quickly.

Fortunately, the rapid temperature increase can be avoided simply by continuing to inject particles into the stratosphere. Indeed, the harms of rapid temperature increase provide strong incentive to not to stop particle injection in the first place. Under normal circumstances, people would have to be either incompetent or malicious to stop particle injection. Assuming the geoengineering is managed by responsible parties, abrupt halt may be unlikely, making stratospheric geoengineering relatively safe. This is under normal circumstances. However, particle injection may nonetheless halt if some other catastrophe occurs, such as a war or a pandemic, that prevents people from continuing the injections. The result would be a "double catastrophe" of rapid temperature increase hitting a population already vulnerable from the first catastrophe [29]. This double catastrophe could be very harmful; potentially it could even result in human extinction. This makes for a rather severe downside.

Figure 1 depicts the double catastrophe scenario. The figure shows average global temperature vs. time for three scenarios: (1) the world without geoengineering, in which temperatures gradually rise due to greenhouse gas emissions; (2) ongoing geoengineering, in which temperatures remain indefinitely at a low, stable level (around 13ºC); and (3) abrupt geoengineering halt (around the year 2080), in which temperatures rapidly rise towards where they would have been without geoengineering. Figure 1 also indicates when an initial catastrophe would occur (shortly before 2080) in a double catastrophe scenario.

The great downside dilemma for stratospheric geoengineering is the dilemma of whether to inject particles into the stratosphere. On one hand, stratospheric geoengineering could lower temperatures, avoiding many harms of global warming. On the other hand, it poses a risk of rapid temperature increase that could even result in human extinction. So, should stratospheric geoengineering be pursued?

A key factor in resolving the dilemma is understanding how bad the impacts of global warming could get without geoengineering. The floods, heat waves, and other effects already being observed will almost certainly get worse. This is bad, but there is an even worse impact potentially on the horizon: the exceedance of mammalian thermal limits. The limits depend on wet bulb temperature, which is a combination of "regular" dry bulb temperature and humidity. When wet bulb temperature goes above 35ºC, mammals—including humans—can no longer perspire to regulate our body temperature, causing us to overheat and die. Currently, wet bulb temperatures never exceed the 35ºC limit, but under some possible global warming scenarios, the limit would sometimes be exceeded in much of the land surface of the planet [30]. Unless humans and other mammals took shelter in air conditioning, they would die. Under these conditions, it may be difficult to keep civilization intact in the warmer regions or even worldwide.

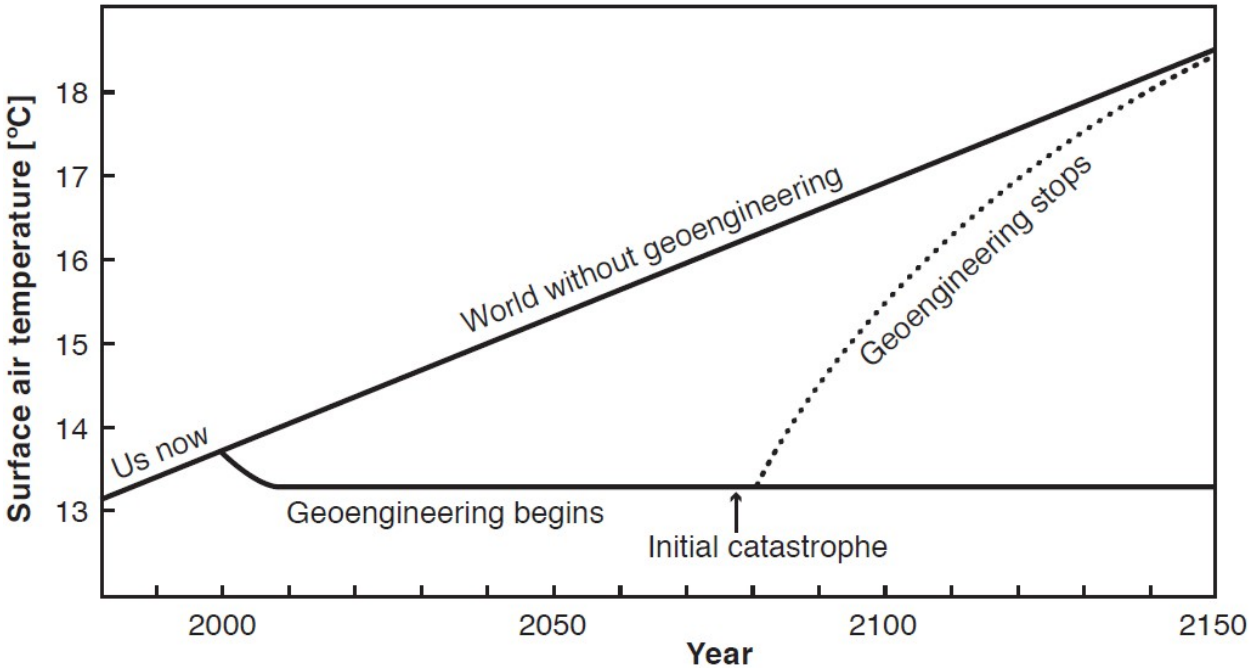

Figure 1: Average global temperature for three scenarios: no geoengineering, ongoing geoengineering, and geoengineering that abruptly stops. The initial catastrophe corresponds to a geoengineering double catastrophe [29].

Another perspective on the potential severity of global warming comes from looking at the long-term co-evolution of the human species and Earth climates. The species *Homo Sapiens Sapiens* is dated at around 200,000 years old. This means that humans have lived through about two full glacial-interglacial cycles, i.e. ice ages and the warm periods between them, which cycle back and forth on time-scales of about 100,000 years [23]. Archaeological evidence suggests that early *Homo Sapiens Sapiens* and their immediate ancestors had cognitive capabilities comparable to those of contemporary humans [31-32]. However, civilization did not take off until the agricultural revolution, which began around 10,000 years ago and occurred in at least seven independent locations within just a few thousand years. The last 10,000 years coincide with the Holocene, a warm interglacial period with a relatively stable climate, suggesting that this climate may have been crucial for the rise of civilization [33]. Meanwhile, global warming threatens to push temperatures to levels significantly outside the range of recent glacial-interglacial cycles, bringing climates that *Homo Sapiens Sapiens* and its immediate ancestors have never seen before [34]. To the extent that certain climates are essential for human civilization, global warming could be devastating.

This sort of long-term perspective is also helpful for understanding stratospheric geoengineering risk. Global warming could last for centuries, millennia, or even longer [34]. This is a very long time to continue injecting particles into the stratosphere. It is also plenty of time for plenty of catastrophes to occur. For example, risk analysis of nuclear war finds about a 0.1% to 1% chance of nuclear war occurring during any given year [35]. Over hundreds or thousands of years, this makes nuclear war virtually certain to occur. Of course, the world could permanently get rid of nuclear weapons, eliminating the risk. But this might not happen, and meanwhile there are other types of catastrophes to worry about. Over the time-scales of global warming, a stratospheric geoengineering double catastrophe may be quite likely.

So, should stratospheric geoengineering be pursued? At this time, I do not believe a good answer to this question exists. There is too much uncertainty about both the consequences of stratospheric geoengineering and the consequences of *not* stratospheric geoengineering, as well as the probabilities of stratospheric geoengineering abrupt halt. Fortunately, the decision does not need to be made just yet. Global warming is not yet so bad that stratospheric geoengineering is worth the risk. But geoengineering decisions may be made soon; research to reduce the uncertainty should proceed now so that wise decisions can be made [36-37].

When the time comes to decide whether to launch stratospheric geoengineering, the right action to take may also be the more difficult action to take. It is quite plausible that civilization could endure the worst harms of regular global warming, but would collapse from the rapid global warming of stratospheric geoengineering abrupt halt. If this is the case, then it would be in civilization's long-term interest to abstain from stratospheric geoengineering and suffer through regular global warming. This abstention may be best regardless of how painful regular global warming could get and regardless of how unlikely stratospheric geoengineering abrupt halt would be. It is just that important to ensure the long-term viability of human civilization. But this is hardly an encouraging prospect, dooming humanity to the pains of regular global warming, when lower temperatures could so easily be produced.

Meanwhile, civilization can lessen the stratospheric geoengineering dilemma by reducing greenhouse gas emissions. Regardless of any broader failures at emissions reductions, every additional bit helps. And reducing emissions helps with both sides of the dilemma: it lessens the severity of both regular global warming and the rapid temperature increase from stratospheric geoengineering abrupt halt. As discussed further below, many options for reducing emissions come with minimal dilemmas of their own, making them excellent options to pursue.

3.2. Artificial General Intelligence

In January 2011, two world-leading players of the game show *Jeopardy!* took on an IBM computer named Watson. Watson won the game convincingly. During the last Final Jeopardy! round, human contestant Ken Jennings wrote below his response question, "I for one welcome our new computer overlords". It was a humorous moment in the long rise of artificial intelligence (AI). But how high can AI rise? Could AI actually become the overlords of humanity, taking over the world? And should such a development be welcomed?

At the outset, it is important to distinguish between two types of AI: narrow and general. Narrow AI is intelligent in specific domains but cannot reason outside the domains it was designed for. Narrow AI is by now ubiquitous across a breadth of contexts, from searching the web to playing games like *Jeopardy!* Narrow AI can be quite useful, and can also pose some risks. But it is not expected to take over the world, because controlling the world requires capabilities across many domains. General AI (AGI) is intelligent across a wide range of domains. Humans are also intelligent across many domains, but this does not mean that AGI would necessarily think like humans do. An AGI may not need to think like humans in order to be capable across many domains.[8]

Early AI researchers boldly predicted that human-level AGI would be achieved by dates long since past, as Crevier [39] and McCorduck [40] chronicle. This grandiose failure of prediction led many modern AI researchers to be skeptical about the prospects of AGI [41]. However, there remains an active AGI research community [42]. Experts in the field diverge widely about when

[8] AGI was discussed in detail in another paper in the series of papers based on the event "Emerging Technologies and the Future of Humanity" [38].

AGI is likely to be achieved and on its impacts if or when it is achieved [43-45]. But some of the predictions about impacts are quite dramatic.

One line of thinking posits that an AGI, or at least certain types of AGIs, could essentially take over the world. This claim depends on two others. First, power benefits from intelligence, such that the most intelligent entities will tend to have the most power. Second, an AGI can gain vastly more intelligence than humans can, especially if the AGI can design an even more intelligent AGI, which designs a still more intelligent AGI, and so on until an "intelligence explosion" [46] or "Singularity" [47-48] occurs. The resulting "superintelligent" AGI [49] could be humanity's final invention [50] because the AGI would then be fully in control. If the AGI is "Friendly" to humanity [51], then it potentially could solve a great many of humanity's problems. Otherwise, the AGI will likely kill everyone inadvertently as it pursues whatever goals it happened to be programmed with—for example, an AGI programmed to excel at chess would kill everyone while converting the planet into a computer that enabled it to calculate better chess moves [52].

Per this line of thinking, an AGI would be much like a magic genie, such as the one depicted in the film *Aladdin* (John Musker and Ron Clements, directors, 1992). The genie is all-powerful but obligated to serve its master. The master can wish for almost anything, but should be careful what he or she wishes for. Indeed, genie stories are often stories of unintended consequences. For example, in the penultimate scene of *Aladdin*, Jafar wishes to become a genie. He was eager to gain the genie's powers, but ended up trapped in servitude (and stuck inside a small lamp). The story with AGI may be similar. The AGI would do exactly what its human programmers instructed it to do, regardless of whether the programmers would, in retrospect, actually want this to happen. In attempting to program the AGI to do something desirable, the programmers could end up dead, along with everyone else on the planet.[9]

If this line of thinking is correct, or even if it has at least some chance of being correct, then AGI poses a great downside dilemma. Should an AGI be built and launched? Given the possibility of being destroyed by AGI, it might appear that AGI should simply not be built. Doing so would ensure that humanity retains control of itself and its fate. But for several reasons, the situation is not so simple.

A first complication is that AGI might be Friendly or otherwise beneficial to humanity, or to the world. The benefits of a Friendly AGI could be immense. Imagine having the perfect genie: unlimited wishes that are interpreted as you intended them to be, or maybe even better than you intended them to be. That could go quite well. Perhaps there would be no more poverty or pollution. Perhaps space colonization could proceed apace. Perhaps the human population could double, or triple, or grow tens, hundreds, or thousands of times larger, all with no decline in quality of life. A Friendly AGI might be able to make these things possible.

Decision-making on AGI should balance this large potential upside with the also-large downside risk. For example, suppose the AGI had a 50% chance of killing everyone and a 50% chance of doubling the human population with no decline in quality of life. The expected population would be equally large with or without the AGI. Does this mean that humanity is indifferent to launching the AGI? If it was a 51% chance of doubling the population, vs. 49% for killing everyone, does this mean humanity would rather launch the AGI? What if it was a chance of the population increasing by a factor of ten, or a thousand? These are important types of questions to answer when making decisions about launching an AGI.

[9] Arguably, such a result would at least be better than an eternity trapped in a small lamp.

A second complication is that AGI is not the only threat that humanity faces. In the absence of AGI, humanity might die out anyway because of nuclear weapons, global warming, or something else. If AGI succeeds, then these other threats could go away, solved by our new computer overlords. That is a significant upside for AGI. What if an AGI has a 50% chance of killing everyone, but absent AGI, humanity has a 60% of dying out from something else? Should the AGI be launched?

The dilemma for AGI can thus look a lot like that for stratospheric geoengineering. Imagine, some years into the future, humanity finds itself in a difficult spot. Perhaps global warming is bringing great harm, and other environmental stressors are as well. Perhaps major countries are on the brink of war with nuclear weapons or something even more destructive. Perhaps poverty is rampant, life unsafe and unpleasant. Perhaps other solutions, including stratospheric geoengineering, are found to be unsafe or otherwise undesirable. And perhaps there is no hope in sight of conditions improving. In this case, taking the risk of launching a AGI could start to look attractive. Indeed, in terms of the long-term success of human civilization, it might even be the right thing to do. Or, the right thing may be to suffer through without AGI. It would depend on the details, just as it would for stratospheric geoengineering or a desperate game of Russian roulette.

Following this logic, one way to help reduce AGI risk is to improve the general human condition. By keeping humanity out of desperate circumstances, the risk of AGI can be made to look less attractive. This opens up a wide range of opportunities to help reduce AGI risk, from reducing greenhouse gas emissions to improving conditions for the world's poor. But the merits of this approach depend on how the AGI would be developed.

The third complication is that AGI development could involve basic computing resources and technologies of growing economic importance. AGI is not like nuclear weapons, which require exotic materials. AGI could be developed on any sufficiently powerful computer. Computing power is steadily growing, a trend known as Moore's Law. Meanwhile, narrow AI is of increasing technological sophistication and economic importance. At the time of this writing, driverless cars are just starting to hit the streets around the world, with promise to grow into a major industry. Differences between narrow AI and AGI notwithstanding, these AI advances may be able to facilitate the development of AGI.

Given the risks of AGI, it may seem attractive or even wise to relinquish precursor hardware and software technologies, potentially including certain narrow AI and the computer systems they run on [53]. But given the pervasiveness of these technologies, it may be difficult to do so. Here lies another dilemma. Would humanity be willing to sacrifice much of its computing technology in order to avoid an AGI catastrophe? Should it?

The dilemma here resembles that faced in the recent film *Transcendence* (Wally Pfister, director, 2014). The film shows an AGI that has been launched and is steadily taking over the world. The AGI is in many ways beneficial or even Friendly, but the humans who are close to it become increasingly skeptical and decide to shut it down. (More precisely, they persuade the AGI to shut itself down, since the AGI was still in control.) However, in shutting it down, humanity had to sacrifice the internet, and potentially also other electronics. A case can be made that the AGI should not have been shut down: without the internet and other electronics, the long-term prospects for human civilization could be severely limited, such that humanity would be better off keeping the AGI intact and hoping for the best [54].

As with stratospheric geoengineering, AGI launch decisions do not need to be made right now. However, for AGI there is great uncertainty about how much time remains. Experts are

sharply divided on how long it will take to achieve AGI, with some doubting that it will ever occur. Given this uncertainty, and the high stakes of the launch decision, it is not at all too early to assess which AGIs should or should not be launched, and to create the conditions that can help ensure better outcomes whether or not an AGI is launched.

4. Technologies Without Great Downside Dilemma

Not all technologies present a great downside dilemma. These technologies may be disruptive, may have downsides, and may carry risks, but they do not threaten catastrophic harm to human civilization. Or, to the extent that they could threaten catastrophic harm, they do not increase the risk of catastrophe beyond what it would be without the technology, or do not increase the risk to any significant extent. Some of these technologies even hold great potential to improve the human condition, including by reducing other catastrophic risks. These latter technologies are especially attractive and in general should be pursued to the extent that their benefits and cost-effectiveness are competitive with other options for improving the human condition (and achieving any other goals). Three such emerging technologies are discussed here.

4.1 Sustainable Design

Sustainable design refers broadly to the design of technologies oriented towards improving the environment, advancing sustainability, and related goals. These technologies promise to reduce the harms of climate change and other environmental problems. Quite a lot of such technologies are already in use, from the humble bicycle to advanced solar technologies. This is a vast technology space, and a lot has been said about these technologies elsewhere [55], so a full review here is unwarranted. What is worth noting here is that these technologies can reduce the risk of environmental catastrophes like climate change, and are often also worth pursuing for other reasons. For example, technology that uses energy, water, and other resources more efficiently can save money by avoiding purchases of these resources. Technologies like bicycles can make people healthier by giving them more exercise. Where sustainable design comes with such co-benefits, it is an especially attractive option. But given the catastrophic potential of environmental risks, some sustainable design, and potentially quite a lot of it, is worth pursuing even if it otherwise comes at an expense.

4.2 Nuclear Fusion Power

Nuclear fusion power is perhaps the Holy Grail of sustainable design. It promises a clean, safe, abundant energy source. If nuclear fusion power can be realized, and if it can be made affordable, then humanity's energy needs could potentially be fully met. And with abundant energy, a lot of other opportunities open up. For example: Ocean water could be desalinated, eliminating water resource scarcities. Carbon dioxide could be removed from the atmosphere, which is another form of geoengineering, and a much safer one at that. Countries can develop their economies without worrying nearly as much about their environmental impact and without worrying about being dependent on another country's energy resources.

One major long-term benefit of fusion power relates back to the fossil fuels it would replace. With fusion power, humanity can keep the rest of the fossil fuels underground, ready and waiting for when they will really be needed. That time will come sometime within the upcoming hundreds of thousands of years, when Earth's climate cycles back to a glacial period: a new ice age. The exact timing of the next glacial period is uncertain, and depends on, among other things, how much greenhouse gas humanity emits [34]. But, barring any other radical changes to the

global Earth system (such as its dismantling by a runaway AGI), a glacial period will eventually occur. And when it does, it could help to still have some fossil fuel around to lessen the bite of the global cooling [56, p.234-235]. This would be yet another form of geoengineering, one with the long-term interests of human civilization in mind.

Unfortunately, it is not clear if or when the Holy Grail of fusion power will be achieved. Fusion power research has been going on for decades [57], and it may take more decades still. With such a long development period, fusion power is a modern analog to cathedrals. Many cathedrals took a century or longer to build. This includes at least one cathedral currently under construction, Sagrada Família in Barcelona, whose construction began in 1882 and has no clear projected completion date. Humanity's track record with cathedrals indicates its capability to complete large, multi-century, intergenerational projects. Perhaps the fusion power project will be completed too.

Unlike cathedrals, it is not known if it is even possible to complete the fusion power project: to make fusion power a major energy source for human civilization. Already, it is possible to generate power from nuclear fusion. First came uncontrolled fusion—fusion bombs—beginning with the detonation of Ivy Mike in 1952. Soon after came controlled fusion, beginning with Scylla 1 at Los Alamos National Laboratory in 1958 [58]. Controlled fusion is what can be used for electricity generation. However, controlled fusion thus far has always consumed more energy than it generates. A major breakthrough recently occurred at the National Ignition Facility at Lawrence Livermore National Laboratory: for the first time, a net energy gain occurred within the fuel that triggers the fusion [59]. However, the fuel is just one part of the fusion process; the National Ignition Facility experiment consumed overall about 100 times more energy than it generates. But, clear progress is being made. On the other hand, much more progress is needed still, and it is not clear if or when net energy gain will be achieved, or if it would be affordable.

If affordable fusion power is achieved, it would be transformative. The fuels are deuterium and lithium, supplies of which can last for thousands to billions of years, depending on power plant design, and there could be no significant radioactive waste [60]. While fusion reactors potentially could be used to generate materials for nuclear weapons, their weapons proliferation risk would be lower, potentially much lower, than it is for fission power [61-62]. While fusion power research is expensive and the prospects for success uncertain, the potential benefits are, in my own estimation, sufficient to justify ongoing investment. This cathedral is well worth attempting to build.

4.3 Space Colonization

The dream of living beyond Earth may be as old as humanity itself. Within the last century, concrete steps have been taken towards this dream. The project of colonizing space may take even longer than the project of fusion power, perhaps orders of magnitude longer. But it comes with its own set of sizable benefits, with relatively little risk. One benefit is the emotional inspiration that humanity can draw from marveling at its cosmic achievement [63]. Other notable benefits are more practical, but no less great.

One major benefit of space colonization is the protection it offers against global catastrophes on Earth. If humanity has self-sufficient space colonies, then it can survive even the complete destruction of its home planet. A spacefaring civilization is a more resilient civilization. This benefit has prompted calls for space colonization [64]. However, space colonization using current technology would be highly expensive and perhaps not even feasible, rendering other options for protecting against catastrophes, including Earth-based refuges, the more cost-

effective option [65-66]. The protections that space colonization could offer do not justify investment in space colonization at this time.

While space colonization can protect against harms, it can also enable major benefits on its own. The opportunities for civilization are, quite literally, astronomically greater beyond Earth than on it. Indeed, the astronomic potential for human civilization is a main reason why great downside dilemmas and other global catastrophic risk decisions are so important to resolve. But again, this does not mean that humanity should invest in space colonization at this time. Instead, it would be wise to focus on the catastrophic threats it faces, such that future generations can go on to colonize space and achieve astronomically great success as a civilization.

5. Conclusion

The fate of human civilization now hangs in the balance. As James Martin put it [9], humanity is going through a turbulent river full of many threats to its survival. Many of these threats derive from risky emerging technologies like stratospheric geoengineering and artificial general intelligence. Some threats also derive from established technologies like nuclear weapons and radio telescopes for messaging to extraterrestrials. And other technologies do not pose a significant threat, including sustainable design technologies, nuclear fusion power, and space colonization. Meanwhile, all of these technologies, if used properly, could help humanity navigate the turbulence. And if the turbulence is successfully navigated, a very long and bright future awaits. Humanity's future could include billions of years on Earth as well as a much bigger and longer existence across the universe. Human civilization and its descendents can achieve many great things, if only it has the opportunity. Navigating the turbulence—preventing civilization-ending global catastrophe—is thus a crucial task for this era of human civilization.

The great downside dilemma for risky emerging technologies could be an especially difficult stretch of turbulence for humanity to navigate. Technologies like stratospheric geoengineering and artificial general intelligence pose great temptations, especially if humanity finds itself in difficult circumstances. For the long-term sake of human civilization, it may be best to abstain from the technologies, but over the short-term, abstention could mean suffer through life without them. Global warming is just one of several forces that could put humanity in desperate circumstances in the not-too-distant future, making risky technologies especially attractive.

If the right decisions are to be made about these various technologies—and that could mean taking the risk of using them—then two things are needed. First, the risks must be understood. People must know what the right decision is. This means characterizing the probabilities that the technologies will fail, the severity of harm if they do fail, and humanity's prospects if the technologies are not used. But, as they say, knowing is only half the battle. The other half is applying the knowledge. The second thing needed is for decision-making procedures to be in place such that bad risks are not taken. Accomplishing this means bringing together the many people involved in risky technology development, from scientists and engineers to government regulators. Some scientists and engineers might not like having their work regulated, but this only underscores the importance of including them in the process, so their concerns can be addressed, as can anyone else's.

Many jurisdictions already regulate a variety of technologies, in light of the risks they pose. This is a good step. But emerging technologies pose new challenges that must be addressed in turn. And the global nature of the worst catastrophes suggests a role for international cooperation [67]. Efforts at smaller scales can also play a role, including the daily actions everyone can make to protect the environment, promote peace, and otherwise keep humanity out of desperate

circumstances. For the sake of human civilization—indeed, for the sake of the universe—actions across all these scales are well worth taking.

Acknowledgments
This paper is based on a talk at Kungliga Vetenskapsakademien (Royal Swedish Academy of Sciences), 17 March 2014; the audience there provided helpful discussion. Suzanne Lidström provided valuable guidance on the design of this paper, and gave detailed editorial comments. Olle Häggström and two anonymous reviewers provided additional comments. Melissa Thomas-Baum provided helpful feedback on an earlier version of this paper and produced the figure. Steven Umbrello provided excellent assistance with manuscript formatting. Any remaining errors or shortcomings are the author's alone.

References

[1] Ng Y-K 1991 Should we be very cautious or extremely cautious on measures that may involve our destruction? *Social Choice and Welfare* **8** 79-88.
[2] Leslie J 1996. *The End of the World: The Science and Ethics of Human Extinction* (London: Routledge)
[3] Tonn B E 2002 Distant futures and the environment. *Futures* **34** 117-132
[4] Rees M 2003 *Our Final Century: Will the Human Race Survive the Twenty-first Century?* (Oxford: William Heinemann)
[5] Posner R 2004 *Catastrophe: Risk and Response* (Oxford: Oxford University Press)
[6] Beckstead N 2013 *On the Overwhelming Importance of Shaping the Far Future*. Doctoral Dissertation, Department of Philosophy, Rutgers University.
[7] Bostrom N 2013 Existential risk prevention as a global priority. *Global Policy* **4(1)** 15-31
[8] Maher T M Jr, Baum S D 2013 Adaptation to and recovery from global catastrophe. *Sustainability* **5(4)** 1461-1479
[9] Martin J 2007 *The Meaning of the 21st Century* (New York: Riverhead Penguin)
[10] Reed B C 2014 The Manhattan Project. *Physica Scripta* **89(10)** 108003
[11] Konopinski E J., Marvin C., Teller E 1946 *Ignition of the Atmosphere with Nuclear Bombs Report LA-602* (NM: Los Alamos Laboratory)
[12] Russell B *et al* 1955 *The Russell-Einstein Manifesto*
[13] Helfand I 2013 *Nuclear Famine: Two Billion People at Risk*. International Physicians for the Prevention of Nuclear War, http://www.ippnw.org/pdf/nuclear-famine-two-billion at-risk-2013.pdf
[14] Xia L, Robock A., Mills M, Stenke A, Helfand I 2013 Global famine after a regional nuclear war. Under review at Earth's Future. http://climate.envsci.rutgers.edu/pdf/NWXIA4.pdf
[15] Baum S D 2014 Reconciling nuclear winter risk with nuclear weapons policy: The search for safer deterrence. Under review at *Contemporary Security Policy.*
[16] Michaud M A G 2007 *Contact with Alien Civilizations: Our Hopes and Fears About Encountering Extraterrestrials* (New York: Copernicus Books)
[17] Baum S D, Haqq-Misra J D, Domagal-Goldman S D 2011 Would contact with extraterrestrials benefit or harm humanity? A scenario analysis. *Acta Astronautica* **68(11)** 2114-2129
[18] Haqq-Misra J, Busch M W, Som S M., Baum S D 2013 The benefits and harm of transmitting into space. *Space Policy* **29(1)** 40-48

[19] Wright J T *et al* 2011 The exoplanet orbit database. *Publications of the Astronomical Society of the Pacific* **123(902)** 412-422
[20] Brin D undated. Shouting at the cosmos: How SETI has taken a worrisome turn into dangerous territory. http://www.davidbrin.com/shouldsetitransmit.html
[21] CNN Wire Staff 2010 Urgent cry for help as death toll rises from Pakistan flooding. *CNN World*, 16 August. http://www.cnn.com/2010/WORLD/asiapcf/08/16/pakistan.floods
[22] Field C *et al* 2014 Technical Summary In *Climate Change 2014*: *Impacts, Adaptation, and Vulnerability*, available at http://ipcc.ch/report/ar5/wg2/
[23] Archer D 2008 *The Long Thaw: How Humans are Changing the Next 100,000 Years of Earth's Climate* (Princeton: Princton University Press)
[24] Mann M E 2012 *The Hockey Stick and the Climate Wars: Dispatches from the Front Lines* (New York: Columbia University Press)
[25] Baum S D, Haqq-Misra J D, Karmosky C 2012 Climate change: Evidence of human causes and arguments for emissions reduction. *Sci. Eng. Ethics* **18(2)** 393-410
[26] Intergovernmental Panel on Climate Change 2014 *Fifth Assessment Report*, available at http://ipcc.ch/report/ar5
[27] Caldeira K, Bala G, Cao L 2013 The science of geoengineering. *Annual Review of Earth and Planetary Sciences* **41** 231-256
[28] Matthews H D, Caldeira K 2007 Transient climate–carbon simulations of planetary geoengineering. *Proceedings of the National Academy of Sciences* **104** 9949-9954
[29] Baum S D, Maher Jr T M, Haqq-Misra J 2013 Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse. *Environment, Systems & Decisions* **33(1)** 168-180
[30] Sherwood S C, Huber M 2010 An adaptability limit to climate change due to heat stress. *Proceedings of the National Academy of Sciences* **107(21)** 9552-9555
[31] McBrearty S, Brooks A S 2000 The revolution that wasn't: a new interpretation of the origin of modern human behavior. *Journal of Human Evolution* **39(5)** 453-563
[32] McBrearty S 2013 Advances in the study of the origin of humanness. *Journal of Anthropological Research* **69(1)** 7-31
[33] Richerson P J, Boyd R, Bettinger R L 2001 Was agriculture impossible during the Pleistocene but mandatory during the Holocene? A climate change hypothesis *American Antiquity* **66(3)** 387-411
[34] Archer D, Ganopolski A 2005 A movable trigger: Fossil fuel CO2 and the onset of the next glaciation. *Geochemistry, Geophysics, Geosystems* **6(5)**
[35] Barrett A M Baum S D., Hostetler K 2013 Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia *Science & Global Security* **21(2)** 106-133
[36] Keith D W, Parson E, Morgan M G 2010 Research on global sun block needed now. *Nature* **463(7280)** 426-427
[37] Keith D 2013 *A Case for Climate Engineering* (MA: MIT Press)
[38] Sotala K, Yampolskiy R V 2014 Responses to catastrophic AGI risk: A survey. *Physica Scripta*, December
[39] Crevier D 1993 *AI: The Tumultuous History of the Search for Artificial Intelligence* (New York: Basic Books)
[40] McCorduck P 2004 *Machines Who Think: 25th Anniversary Edition*. (MA: A.K. Peters)

[41] Horvitz E, Selman B 2009 *Interim Report from the Panel Chairs, AAAI Presidential Panel on Long-Term AI Futures*. http://www.aaai.org/Organization/Panel/panel-note.pdf
[42] Goertzel B, Pennachin C ed 2007 *Artificial General Intelligence* (New York: Springer Verlag)
[43] Baum S D, Goertzel B, Goertzel T G 2011 How long until human-level AI? Results from an expert assessment. *Technological Forecasting & Social Change*. **78(1)** 185-195
[44] Armstrong S, Sotala K 2012 *Beyond AI: Artificial Dreams* ed Ircing P, Zackova E, Polak M, Schuster R. (Pilsen: University of West Bohemia)
[45] Eden A H, Moor J H, Soraker J H., Steinhart E 2013 *Singularity Hypotheses: A Scientific and Philosophical Assessment* (Springer)
[46] Good I J 1965 Speculations Concerning the First Ultraintelligent Machine. In ed Alt F L, Rubinoff M *Advances in Computers, Advances in Computers*. Academic Press **6** 31–88
[47] Kurzweil R 2005 *The Singularity Is Near: When Humans Transcend Biology* (New York: Penguin Group)
[48] Chalmers D 2010 The Singularity: A philosophical analysis. *Journal of Consciousness Studies* **17** 7-65
[49] Bostrom N 2014 *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press)
[50] Barrat J 2013 *Our Final Invention: Artificial Intelligence and the End of the Human Era* (Macmillan)
[51] Yudkowsky E 2011 Complex value systems in Friendly AI. In ed Schmidhuber J., Thórisson K R., Looks M *Artificial General Intelligence: 4th International Conference Proceedings* (Berlin: Springer) 388-393
[52] Omohundro S 2014 Autonomous technology and the greater human good. *Journal of Experimental & Theoretical Artificial Intelligence* **26(3)** 303-315
[53] Joy B 2000 Why the future doesn't need us. *Wired* **8(04)** 238-262 (http://archive.wired.com/wired/archive/8.04/joy.html)
[54] Baum S D 2014 Film review: Transcendence. *Journal of Evolution and Technology* **24(2)** 79-84
[55] Edenhofer O *et al* 2014. Technical summary. In *Climate Change 2014: Mitigation of Climate Change*, available at http://ipcc.ch/report/ar5/wg3
[56] Stager C 2011 *Deep Future: The Next 100,000 Years of Life on Earth* (New York: Thomas Dunne Books)
[57] Lehnert B 2013 Half a century of fusion research towards ITER. *Physica Scripta* **87(1)** 018201
[58] Phillips J A 1983 Magnetic fusion. *Los Alamos Science* **(7)** 64-67
[59] Hurricane O A *et al* 2014 Fuel gain exceeding unity in an inertially confined fusion implosion. *Nature* **506** 343-348
[60] Ongena J, Van Oost G 2002 Energy for future centuries. *Transactions of Fusion Science and Technology* **41** 3-14
[61] Glaser A, Goldston R J 2012 Proliferation risks of magnetic fusion energy: clandestine production, covert production and breakout. *Nuclear Fusion* **52(4)** 043004
[62] Franceschini G., Englert M., Liebert W 2013 Nuclear fusion power for weapons purposes: An exercise in nuclear proliferation forecasting. *Nonproliferation Review* **20(3)** 525-544
[63] Avdeyev S *et al* 2011 Human space exploration–A global trans-cultural quest. *Space Policy* **27(1)** 24-26

[64] BBC News 2006 Move to new planet, says Hawking. 30 November http://news.bbc.co.uk/1/hi/uk/6158855.stm
[65] Sandberg A, Matheny J G, Ćirković M M 2008 How can we reduce the risk of human extinction? *Bulletin of the Atomic Scientists* http://thebulletin.org/how-can-we-reduce-risk-human-extinction
[66] Baum S D 2009 Cost–benefit analysis of space exploration: Some ethical considerations. *Space Policy* **25(2)** 75-80
[67] Wilson G S 2013 Minimizing global catastrophic and existential risks from emerging technologies through international law. *Virginia Environmental Law Journal* **31(2)** 307-364