EXISTENTIAL RISK PREVENTION AS GLOBAL PRIORITY

(2012) Nick Bostrom

Faculty of Philosophy & Oxford Martin School

University of Oxford

www.nickbostrom.com

www.existential-risk.org

[*Global Policy* (2013), forthcoming]

ABSTRACT

Existential risks are those that threaten the entire future of humanity. Many theories of value imply that even relatively small reductions in net existential risk have enormous expected value. Despite their importance, issues surrounding human-extinction risks and related hazards remain poorly understood. In this paper, I clarify the concept of existential risk and develop an improved classification scheme. I discuss the relation between existential risks and basic issues in axiology, and show how existential risk reduction (via the maxipok rule) can serve as a strongly action-guiding principle for utilitarian concerns. I also show how the notion of existential risk suggests a new way of thinking about the ideal of sustainability.

KEYWORDS: existential risk, catastrophic risk, future of humanity, human extinction, sustainability, maxipok, population ethics

1. The maxipok rule

1.1. Existential risk and uncertainty

An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development (Bostrom 2002). Although it is often difficult to assess the probability of existential risks, there are many reasons to suppose that the total such risk confronting humanity over the next few centuries is

significant. Estimates of 10-20% total existential risk in this century are fairly typical among those who have examined the issue, though inevitably such estimates rely heavily on subjective judgment.[1] The most reasonable estimate might be substantially higher or lower. But perhaps the strongest reason for judging the total existential risk within the next few centuries to be significant is the extreme magnitude of the values at stake. Even a small probability of existential catastrophe could be highly practically significant (Bostrom 2003; Matheny 2007; Posner 2004; Weitzman 2009).

Humanity has survived what we might call *natural existential risks* for hundreds of thousands of years; thus it is prima facie unlikely that any of them will do us in within the next hundred.[2] This conclusion is buttressed when we analyze specific risks from nature, such as asteroid impacts, supervolcanic eruptions, earthquakes, gamma-ray bursts, and so forth: Empirical impact distributions and scientific models suggest that the likelihood of extinction because of these kinds of risk is extremely small on a time scale of a century or so.[3]

In contrast, our species is introducing entirely new kinds of existential risk—threats we have no track record of surviving. Our longevity as a species therefore offers no strong prior grounds for confident optimism. Consideration of specific existential-risk scenarios bears out the suspicion that the great bulk of existential risk in the foreseeable future consists of *anthropogenic existential risks*—that is, those arising from human activity. In particular, most of the biggest existential risks seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology. As our powers expand, so will the scale of their potential consequences—intended and unintended, positive and negative. For example, there appear to be significant existential risks in some of the advanced forms of biotechnology, molecular nanotechnology, and machine intelligence that might be developed in the

---

[1] One informal poll among mainly academic experts on various global catastrophic risks gave a median estimate of 19% probability that the human species will go extinct before the end of this century (Sandberg and Bostrom 2008). These respondents' views are not necessarily representative of the wider expert community. The U.K.'s influential Stern Review on the Economics of Climate Change (2006) used an extinction probability of 0.1% per year in calculating an effective discount rate. This is equivalent to assuming a 9.5% risk of human extinction within the next hundred years (UK Treasury 2006, Chapter 2, Technical Appendix, p. 47).

[2] The strength of this consideration is to some extent blunted by the possibility of observation selection effects casting an "anthropic shadow" on available evidence (Cirkovic, Sandberg and Bostrom 2010).

[3] Cf. Smil 2008.

decades ahead. The bulk of existential risk over the next century may thus reside in rather speculative scenarios to which we cannot assign precise probabilities through any rigorous statistical or scientific method. But the fact that the probability of some risk is difficult to quantify does not imply that the risk is negligible.

Probability can be understood in different senses. Most relevant here is the epistemic sense in which probability is construed as (something like) the credence that an ideally reasonable observer should assign to the risk's materializing based on currently available evidence.[4] If something cannot presently be known to be objectively safe, it is risky at least in the subjective sense relevant to decision making. An empty cave is unsafe in just this sense if you cannot tell whether or not it is home to a hungry lion. It would be rational for you to avoid the cave if you reasonably judge that the expected harm of entry outweighs the expected benefit.

The uncertainty and error-proneness of our first-order assessments of risk is itself something we must factor into our all-things-considered probability assignments. This factor often *dominates* in low-probability, high-consequence risks—especially those involving poorly understood natural phenomena, complex social dynamics, or new technology, or that are difficult to assess for other reasons. Suppose that some scientific analysis A indicates that some catastrophe X has an *extremely* small probability $P(X)$ of occurring. Then the probability that A has some hidden crucial flaw may easily be much greater than $P(X)$.[5] Furthermore, the *conditional* probability of X given that A is crucially flawed, $P(X \mid \neg A)$, may be fairly high. We may then find that most of the risk of X resides in the uncertainty of our scientific assessment that $P(X)$ was small (figure 1) (Ord, Hillerbrand and Sandberg 2010).

[4] Probability is thus indexed to time. Quantities that depend on probability, such as the seriousness of a risk, can vary over time as new information becomes available.

[5] There is ample historical evidence that apparently sound scientific analyses are sometimes crucially flawed.

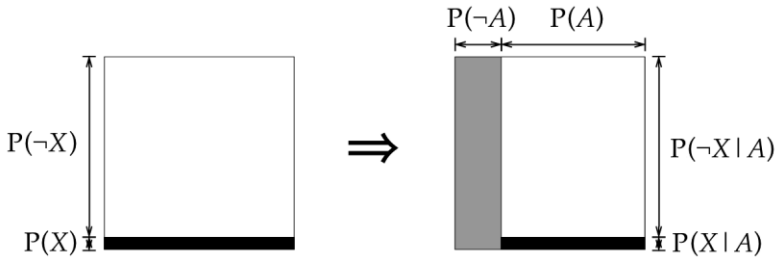

Figure 1: Meta-level uncertainty. Factoring in the fallibility of our first-order risk assessments can amplify the probability of risks assessed to be extremely small. An initial analysis (left side) gives a small probability of a disaster (black stripe). But the analysis could be wrong; this is represented by the gray area (right side). Most of the all-things-considered risk may lie in the gray area rather than in the black stripe.

1.2. Qualitative risk categories

Since a risk is a prospect that is negatively evaluated, the seriousness of a risk—indeed, what is to be regarded as risky at all—depends on an evaluation. Before we can determine the seriousness of a risk, we must specify a standard of evaluation by which the negative value of a particular possible loss scenario is measured. There are several types of such evaluation standard. For example, one could use a utility function that represents some particular agent's preferences over various outcomes. This might be appropriate when one's duty is to give decision support to a particular decision maker. But here we will consider a *normative evaluation*, an ethically warranted assignment of value to various possible outcomes. This type of evaluation is more relevant when we are inquiring into what our society's (or our own individual) risk-mitigation priorities *ought* to be.

There are conflicting theories in moral philosophy about which normative evaluations are correct. I will not here attempt to adjudicate any foundational axiological disagreement. Instead, let us consider a simplified version of one important class of normative theories. Let us suppose that the lives of persons usually have some significant positive value and that this value is aggregative (in the sense that the value of two similar lives is twice that of one life). Let us also assume that, holding the quality and duration of a life constant, its value does not depend on when it occurs or on whether it already exists or is yet to be brought into existence as a result of future events and choices. These assumptions could be relaxed and complications could be introduced, but we will confine our discussion to the simplest case.

Within this framework, then, we can roughly characterize a risk's seriousness using three variables: *scope* (the size of the population at risk), *severity* (how badly this population would be affected), and *probability* (how likely the disaster is to occur, according to the most reasonable judgment, given currently available evidence). Using the first two of these variables, we can construct a qualitative diagram of different types of risk (figure 2). (The probability dimension could be displayed along the z-axis.)

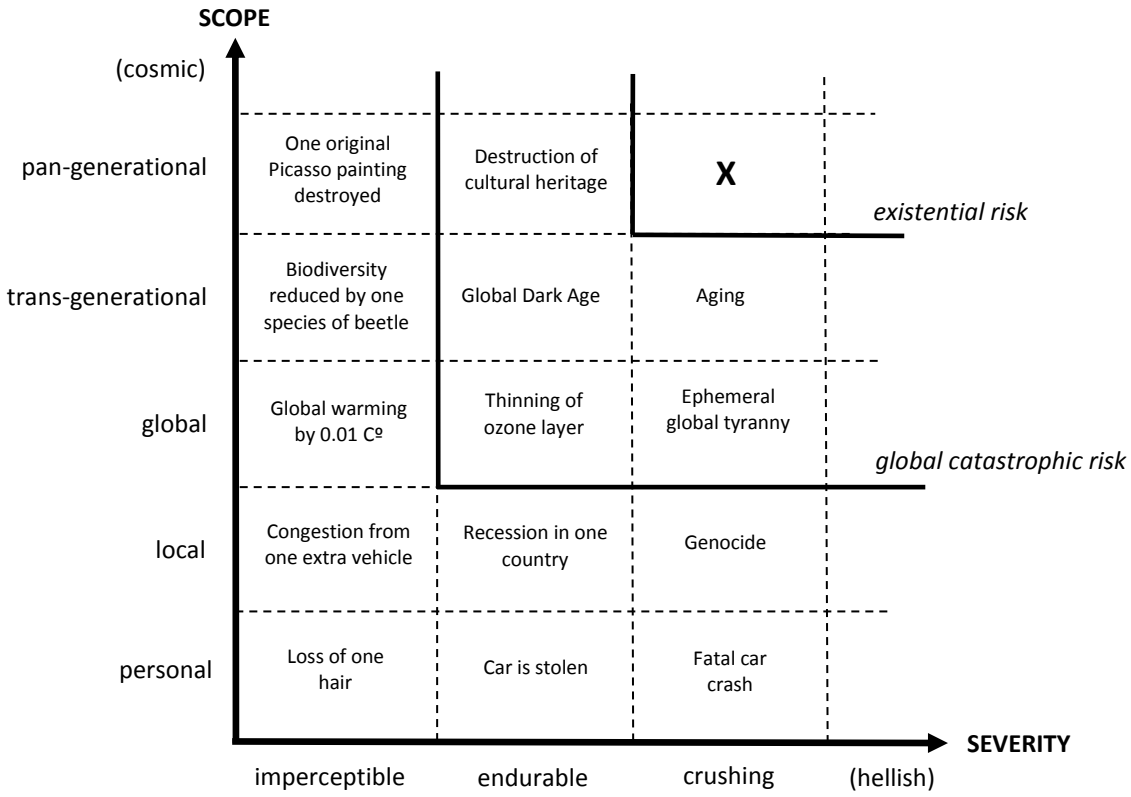

Figure 2: Qualitative risk categories. The scope of a risk can be *personal* (affecting only one person), *local* (affecting some geographical region or a distinct group), *global* (affecting the entire human population or a large part thereof), *trans-generational* (affecting humanity for numerous generations, or *pan-generational* (affecting humanity over all, or almost all, future generations). The severity of a risk can be classified as *imperceptible* (barely noticeable), *endurable* (causing significant harm but not completely ruining quality of life), or *crushing* (causing death or a permanent and drastic reduction of quality of life).

The area marked "X" in figure 2 represents existential risks. This is the category of risks that have (at least) crushing severity and (at least) pan-generational scope.[6] As noted, an existential risk is one that threatens to cause the extinction of Earth-originating intelligent life or the permanent and drastic failure of that life to realize its potential for desirable development. In other words, an existential risk jeopardizes the entire future of humankind.

1.3. Magnitude of expected loss in existential catastrophe

Holding probability constant, risks become more serious as we move toward the upper-right region of figure 2. For any fixed probability, existential risks are thus more serious than other risk categories. But just *how much* more serious might not be intuitively obvious. One might think we could get a grip on how bad an existential catastrophe would be by considering some of the worst historical disasters we can think of—such as the two world wars, the Spanish flu pandemic, or the Holocaust—and then imagining something just a bit worse. Yet if we look at global population statistics over time, we find that these horrible events of the past century fail to register (figure 3).

[6] As indicated in the figure, the axes can be extended to encompass conceptually possible risks that are even more extreme. In particular, pan-generational risks can contain a subclass of risks so destructive that their realization would not only affect or pre-empt future human generations but would also destroy the potential of the part of the universe that lies in our future light cone to produce intelligent or self-aware beings (*cosmic* scope). Further, according to some theories of value there can be states of being that are much worse than nonexistence or death (e.g., horrible incurable diseases), so one could in principle extend the x-axis as well (*hellish* severity). We will not explore these conceptual possibilities in this paper.

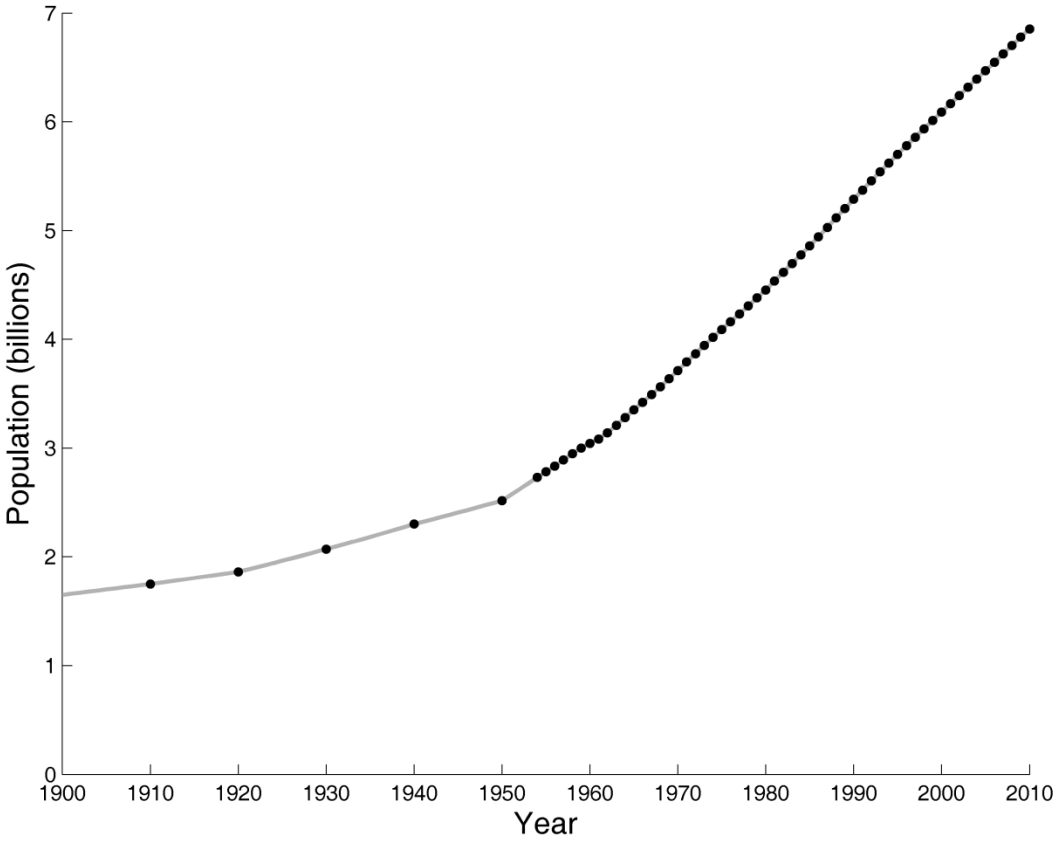

Figure 3: World population over the last century. Calamities such as the Spanish flu pandemic, the two world wars, and the Holocaust scarcely register. (If one stares hard at the graph, one can perhaps just barely make out a slight temporary reduction in the rate of growth of the world population during these events.)

But even this reflection fails to bring out the seriousness of existential risk. What makes existential catastrophes especially bad is not that they would show up robustly on a plot like the one in figure 3, causing a precipitous drop in world population or average quality of life. Instead, their significance lies primarily in the fact that they would destroy the future. The philosopher Derek Parfit made a similar point with the following thought experiment:

> I believe that if we destroy mankind, as we now can, this outcome will be *much* worse than most people think. Compare three outcomes:
> (1) Peace.
> (2) A nuclear war that kills 99% of the world's existing population.
> (3) A nuclear war that kills 100%.

> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is *very much* greater. … The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second. (Parfit 1984, pp. 453-454).

To calculate the loss associated with an existential catastrophe, we must consider how much value would come to exist in its absence. It turns out that the ultimate potential for Earth-originating intelligent life is literally astronomical.

One gets a large number even if one confines one's consideration to the potential for biological human beings living on Earth. If we suppose with Parfit that our planet will remain habitable for at least another billion years, and we assume that at least one billion people could live on it sustainably, then the potential exist for at least 10^{16} human lives of normal duration. These lives could also be considerably better than the average contemporary human life, which is so often marred by disease, poverty, injustice, and various biological limitations that could be partly overcome through continuing technological and moral progress.

However, the relevant figure is not how many people could live on Earth but how many descendants we could have in total. One lower bound of the number of biological human life-years in the future accessible universe (based on current cosmological estimates) is 10^{34} years.[7] Another estimate, which assumes that future minds will be mainly implemented in computational hardware instead of biological neuronal wetware, produces a lower bound of 10^{54} human-brain-emulation subjective life-years (or 10^{71} basic computational operations) (Bostrom 2003).[8] If we make the less

---

[7] This is based on an accelerating universe with a maximal reachable co-moving distance of 4.74 Gpc, a baryonic matter density of $4.55 \cdot 10^{-28}$ kg/m^3, a luminosity ratio of stars ~100, and 1 planet per 1,000 stars being habitable by 1 billion humans for 1 billion years (Gott et al 2005; Heyl 2005). Obviously the values of the last three parameters are debatable, but the astronomical size of the conclusion is little affected by a few orders-of-magnitude change.

[8] This uses an estimate by the late futurist Robert Bradbury that a star can power 10^{42} operations per second using efficient computers built with advanced nanotechnology. Further, it assumes (along with the

conservative assumption that future civilizations could eventually press close to the absolute bounds of known physics (using some as yet unimagined technology), we get radically higher estimates of the amount of computation and memory storage that is achievable and thus of the number of years of subjective experience that could be realized.[9]

Even if we use the most conservative of these estimates, which entirely ignores the possibility of space colonization and software minds, we find that the expected loss of an existential catastrophe is greater than the value of 10^{16} human lives. This implies that the expected value of reducing existential risk by a mere *one millionth of one percentage point* is at least a hundred times the value of a million human lives. The more technologically comprehensive estimate of 10^{54} human-brain-emulation subjective life-years (or 10^{52} lives of ordinary length) makes the same point even more starkly. Even if we give this allegedly lower bound on the cumulative output potential of a technologically mature civilization a mere 1% chance of being correct, we find that the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives.

One might consequently argue that even the tiniest reduction of existential risk has an expected value greater than that of the definite provision of any "ordinary" good, such as the direct benefit of saving 1 billion lives. And, further, that the absolute value of the *indirect* effect of saving 1 billion lives on the total cumulative amount of existential risk—positive or negative—is almost certainly larger than the positive value of the direct benefit of such an action.[10]

---

cosmological estimates mentioned in the previous footnote) that the human brain has a processing power of 10^{17} operations per second and that stars on average last 5 billion years. It does not assume any new star formation. See also (Cirkovic 2004).

[9] For example, if all mass-energy in the accessible universe is saved until the cosmic microwave background temperature ceases to decline (due to the constant horizon temperature of 10^{-29} K) and is then used for computation, this would allow up to 10^{121} thermodynamically irreversible computations (Krauss and Starkman 2000). See also (Cirkovic and Radujkov 2001).

[10] We should stress, however, that there are important unresolved issues in aggregative consequentialism—in particular, in relation to infinite values and extremely small chances (Bostrom 2003 and 2009). We will not discuss these issues here, but in section 5 we will discuss the normative status of the concept of existential risk from some other perspectives.

1.4. Maxipok

These considerations suggest that the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole. It may be useful to adopt the following rule of thumb for such impersonal moral action:

> *Maxipok*
> Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe.

At best, maxipok is a rule of thumb or a *prima facie* suggestion. It is not a principle of absolute validity, since there clearly are moral ends other than the prevention of existential catastrophe. The principle's usefulness is as an aid to prioritization. Unrestricted altruism is not so common that we can afford to fritter it away on a plethora of feel-good projects of suboptimal efficacy. If benefiting humanity by increasing existential safety achieves expected good on a scale many orders of magnitude greater than that of alternative contributions, we would do well to focus on this most efficient philanthropy.

Note that maxipok differs from the popular maximin principle ("Choose the action that has the best worst-case outcome").[11] Since we cannot completely eliminate existential risk—at any moment, we might be tossed into the dustbin of cosmic history by the advancing front of a vacuum phase transition triggered in some remote galaxy a billion years ago—the use of maximin in the present context would entail choosing the action that has the greatest benefit under the assumption of impending extinction. Maximin thus implies that we ought all to start partying as if there were no tomorrow. That implication, while perhaps tempting, is implausible.

[11] Following John Rawls, the term "maximin" is used in a different sense in welfare economics, to denote the principle that (given certain constraints) we ought to opt for the state that maximizes the expectation of the worst-off classes (Rawls 1971). This version of the principle is not necessarily affected by the remarks in the text.

2. CLASSIFICATION OF EXISTENTIAL RISK

To bring attention to the full spectrum of existential risk, we can distinguish four classes of such risk: *human extinction*, *permanent stagnation*, *flawed realization*, and *subsequent ruination*. We define these as follows:

CLASSES OF EXISTENTIAL RISK	
Human extinction	Humanity goes extinct prematurely, i.e., before reaching technological maturity.[12]
Permanent stagnation	Humanity survives but never reaches technological maturity. Subclasses: *unrecovered collapse*, *plateauing*, *recurrent collapse*
Flawed realization	Humanity reaches technological maturity but in a way that is dismally and irremediably flawed. Subclasses: *unconsummated realization*, *ephemeral realization*
Subsequent ruination	Humanity reaches technological maturity in a way that gives good future prospects, yet subsequent developments cause the permanent ruination of those prospects.

By "humanity" we here mean Earth-originating intelligent life and by "technological maturity" we mean the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved.

2.1. Human extinction

Although it is conceivable that, in the billion or so years during which Earth might remain habitable before being overheated by the expanding sun, a new intelligent species would evolve on our planet to fill the niche vacated by an extinct humanity, this is very far from certain to happen. The

[12] One can refer to this more precisely as "early" or "premature" human extinction. Note that humanity can go extinct without instantiating this category if humanity achieves its capability potential and then goes extinct.

probability of a recrudescence of intelligent life is reduced if the catastrophe causing the extinction of the human species also exterminated the great apes and our other close relatives, as would occur in many (though not all) human-extinction scenarios. Furthermore, even if another intelligent species were to evolve to take our place, there is no guarantee that the successor species would sufficiently instantiate qualities that we have reason to value. Intelligence may be necessary for the realization of our future potential for desirable development, but it is not sufficient. All scenarios involving the premature extinction of humanity will be counted as existential catastrophes, even though some such scenarios may, according to some theories of value, be relatively benign. It is not part of the *definition* of existential catastrophe that it is all-things-considered bad, although that will probably be a reasonable supposition in most cases.

Above, we defined "humanity" as Earth-originating intelligent life rather than as the particular biologically defined species *Homo sapiens*.[13] The reason for focusing the notion of existential risk on this broader concept is that there is no reason to suppose that the biological species concept tracks what we have reason to value. If our species were to evolve, or use technology to self-modify, to such an extent that it no longer satisfied the biological criteria for species identity (such as interbreedability) with contemporary *Homo sapiens*, this need not be in any sense a catastrophe. Depending on what we changed into, such a transformation might well be very desirable. Indeed, the permanent foreclosure of any possibility of this kind of transformative change of human biological nature may itself constitute an existential catastrophe.

Most discussion of existential risk to date has focused exclusively on the first of the four classes, "human extinction." The present framework calls attention to three other failure modes for humanity. Like extinction, these other failure modes would involve pan-generational crushing. They are therefore of comparable seriousness, entailing potentially similarly enormous losses of expected value.

[13] We may here take "intelligent" to mean capable of developing language, science, technology, and cumulative culture.

2.2. Permanent stagnation

Permanent stagnation is instantiated if humanity survives but never reaches technological maturity—that is, the attainment of capabilities affording a level of economic productivity and control over nature that is close to the maximum that could feasibly be achieved (in the fullness of time and in the absence of catastrophic defeaters). For instance, a technologically mature civilization could (presumably) engage in large-scale space colonization through the use of automated self-replicating "von Neumann probes." (Freitas 1980; Moravec 1988; Tipler 1980) It would also be able to modify and enhance human biology—say, through the use of advanced biotechnology or molecular nanotechnology (Freitas 1999 and 2003). Further, it could construct extremely powerful computational hardware and use it to create whole-brain emulations and entirely artificial types of sentient, superintelligent minds (Sandberg and Bostrom 2008). It might have many additional capabilities, some of which may not be fully imaginable from our current vantage point.[14]

The permanent destruction of humanity's opportunity to attain technological maturity is a prima facie enormous loss, because the capabilities of a technologically mature civilization could be used to produce outcomes that would plausibly be of great value, such as astronomical numbers of extremely long and fulfilling lives. More specifically, mature technology would enable a far more efficient use of basic natural resources (such as matter, energy, space, time, and negentropy) for the creation of value than is possible with less advanced technology. And mature technology would allow the harvesting (through space colonization) of far more of these resources than is possible with technology whose reach is limited to Earth and its immediate neighborhood.

We can distinguish various kinds of permanent stagnation scenarios: *unrecovered collapse*—much of our current economic and technological capabilities are lost and never recovered; *plateauing*—progress flattens out at a level perhaps somewhat higher than the present level but far

[14] It is not required that a technologically mature civilization *actually deploy* all of these technologies; it is sufficient that they be *available* to it, in the sense that the civilization could easily and quickly develop and deploy them should it decide to do so. Thus, a sufficiently powerful superintelligent-machine civilization that could rapidly invent and implement these and other relevant technologies would already count as technologically mature.

below technological maturity; and *recurrent collapse*—a never-ending cycle of collapse followed by recovery (Bostrom 2009).[15]

The relative plausibility of these scenarios depends on various factors. One might expect that even if global civilization were to undergo a complete collapse, perhaps following a global thermonuclear war, it would eventually be rebuilt. In order to have a plausible permanent collapse scenario, one would therefore need an account of why recovery would not occur.[16] Regarding plateauing, modern trends of rapid social and technological change make such a threat appear less imminent; yet scenarios could be concocted in which, for example, a stable global regime blocks further technological change.[17] As for recurrent-collapse scenarios, they seem to require the postulation of a special kind of cause: one that (*a*) is strong enough to bring about the total collapse of global civilization yet (*b*) is not strong enough to cause human extinction, and that (*c*) can plausibly recur each time civilization is rebuilt to a certain level, despite any random variation in initial conditions and any attempts by successive civilizations to learn from their predecessors' failures. The probability of remaining on a recurring-collapse trajectory diminishes with the number of cycles postulated. The longer the time horizon considered (and this applies also to plateauing) the greater the likelihood that the pattern will be ruptured, resulting in either a breakout in the upward direction toward technological maturity or in the downward direction toward unrecovered collapse and perhaps extinction (figure 4).[18]

---

[15] Not strictly *never*-ending, of course, but a sequence of cycles that goes on for a very long time and ends with human extinction without technological maturity having ever been attained.

[16] An unrecovered collapse scenario might postulate that some critical resource for recovery is permanently destroyed, or that the human gene pool irreversibly degenerates, or perhaps that some discovery is made that enables tiny groups to cause such immense destruction that they can bring down civilization and that the knowledge of this discovery cannot be eradicated.

[17] Improved governance techniques, such as ubiquitous surveillance and neurochemical manipulation, might cement such a regime's hold on power to the extent of making its overthrow impossible.

[18] Another difficulty for the recurring-collapse hypothesis is to account for the fact that we are in the first technological cycle here on Earth. If it is common for there to be many cycles of collapse and recovery (with similar population sizes) then why do we find ourselves in cycle #1? This kind of anthropic consideration might suggest that extinction or transformation is more likely than one would naively suppose.

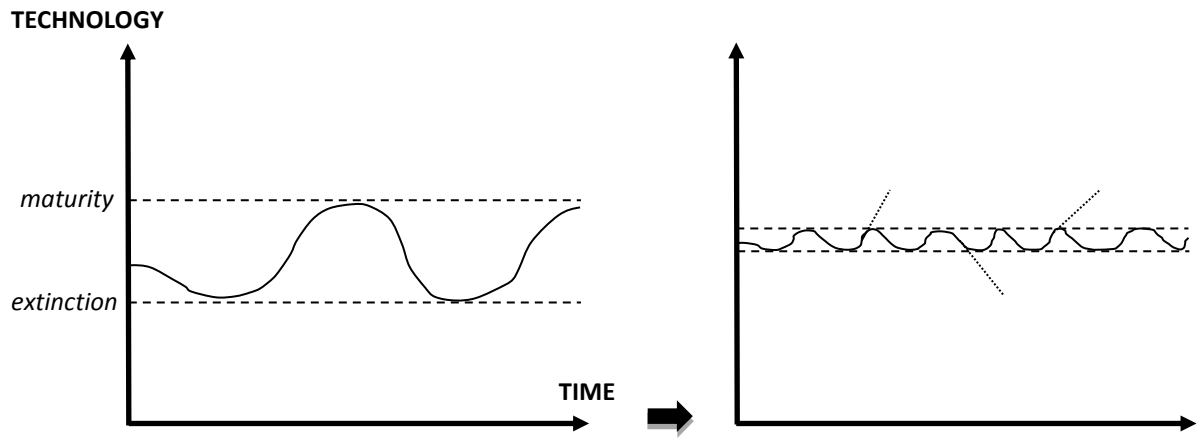

Figure 4: Collapse recurring indefinitely? The modern human condition represents a narrow range of the space of possibilities. The longer the time scale considered, the lower the probability that humanity's level of technological development will remain confined within the interval defined at the lower end by whatever technological capability is necessary for survival and at the upper end by technological maturity.

2.3. Flawed realization

A flawed realization occurs if humanity reaches technological maturity in a way that is dismally and irremediably flawed. By "irremediably" we mean that it cannot feasibly be subsequently put right. By "dismally" we mean that it enables the realization of but a small part of the value that could otherwise have been realized. Classifying a scenario as an instance of flawed realization requires a value judgment. We return to this normative issue in the next section.

We can distinguish two versions of flawed realization: *unconsummated realization* and *ephemeral realization*.

In unconsummated realization, humanity develops mature technology but fails to put it to good use, so that the amount of value realized is but a small fraction of what could have been achieved. An example of this kind is a scenario in which machine intelligence replaces biological intelligence but the machines are constructed in such a way that they lack consciousness (in the sense of phenomenal experience) (Bostrom 2004). The future might then be very wealthy and capable, yet in a relevant sense uninhabited: There would (arguably) be no morally relevant beings there to enjoy the wealth. Even if consciousness did not altogether vanish, there might be a lot less of it than would have resulted from a more optimal use of resources. Alternatively, there might be a

vast quantity of experience but of much lower quality than ought to have been the case: minds that are far less happy than they could have been. Or, again, there might be vast numbers of very happy minds but some other crucial ingredient of a maximally valuable future missing.

In ephemeral realization, humanity develops mature technology that is initially put to good use. But the technological maturity is attained in such a way that the initially excellent state is unsustainable and is doomed to degenerate. There is a flash of value, followed by perpetual dusk or darkness. One way in which ephemeral realization could result is if there are fractures in the initial state of technological maturity that are bound to lead to a splintering of humanity into competing factions. It might be impossible to reintegrate humanity after such a splintering occurred, and the process of attaining technological maturity might have presented the last and best chance for humanity to form a singleton (Bostrom 2006). Absent global coordination, various processes might degrade humanity's long-term potential. One such process is war between major powers, although it is perhaps unlikely that such warring would be never-ending (rather than being eventually terminated once and for all by treaty or conquest).[19] Another such erosive process involves undesirable forms of evolutionary and economic competition in a large ecology of machine intelligences (Hanson 1994). Yet another such process is a space-colonization race in which replicators might burn up cosmic resources in a wasteful effort to beat out the competition (Hanson 1998).

2.4. Subsequent ruination

For completeness, we register a fourth class of existential risks: subsequent ruination. In scenarios of this kind, humanity reaches technological maturity with a "good" (in the sense of being not dismally and irremediably flawed) initial setup, yet subsequent developments nonetheless lead to the permanent ruination of our prospects.

From a practical perspective, we need not worry about subsequent ruination. What happens after humanity reaches technological maturity is not something we can now affect, *other* than by making sure that humanity does reach it and in a way that offers the best possible prospects

[19] Even the threat of a war that never erupts could result in much waste, in terms of expenditures on arms and foregone opportunities for collaboration.

for subsequent development—that is, by avoiding the three other classes of existential risk. Nonetheless, the concept of subsequent ruination is relevant to us in various ways. For instance, in order to estimate how much expected value is gained by reducing other existential risks by a certain amount, we need to estimate the expected value conditional on avoiding the first three sets of existential risks, which requires estimating the probability of subsequent ruination.

The probability of subsequent ruination might be low—and is perhaps extremely low conditional on getting the setup right. One reason is that once we have created many self-sustaining space colonies, any disaster confined to a single planet cannot eliminate all of humanity. Another reason is that once technological maturity is safely reached, there are fewer potentially dangerous technologies left to be discovered. A third reason is that a technologically mature civilization would be superintelligent (or have access to the advice of superintelligent artificial entities) and thus better able to foresee danger and devise plans to minimize existential risk. While foresight will not reduce risk if no effective action is available, a civilization with mature technology can take action against a great range of existential risks. Furthermore, if it turns out that attaining technological maturity without attaining singletonhood condemns a civilization to irreversible degeneration, then if flawed realization is avoided we can assume that our technologically mature civilization can solve global-coordination problems, which increases its ability to take effective action to prevent subsequent ruination.

The main source of subsequent-ruination risk might well be an encounter with intelligent external adversaries, such as intelligent extraterrestrials or simulators. Note, however, that scenarios in which humanity eventually goes extinct as a result of hard physical limits, such as the heat death of the universe, do not count as subsequent ruination, provided that before its demise humanity has managed to realize a reasonably large part of its potential for desirable development. Such scenarios are not existential catastrophes but rather existential successes.

3. CAPABILITY AND VALUE

Some further remarks will help clarify the links between capability, value, and existential risk.

3.1. Convertibility of resources into value

Because humanity's future is potentially astronomically long, the integral of losses associated with persistent inefficiencies is very large. This is why flawed-realization and subsequent-ruination scenarios constitute existential catastrophes even though they do not necessarily involve extinction.[20] It might be well worth a temporary dip in short-term welfare to secure a slightly more efficient long-term realization of humanity's potential.

To avoid flawed realization, it is more important to focus on maximizing long-term efficiency than on maximizing the initial output of value in the period immediately following technological maturation. This is because the quantity of value-structure that can be produced at a given time depends not only on the level of technology but also on the physical resources and other forms of capital available at that time. In economics parlance, humanity's production-possibility frontier (representing the various possible combinations of outputs that could be produced by the global economy) depends not only on the global production function (or "meta-production function") but also on the total amount of all factors of production (labor, land, physical capital goods, etc.) that are available at some point in time. With mature technology, most factors of production are interchangeable and ultimately reducible to basic physical resources, but the amount of free energy available to a civilization imposes hard limits on what it can produce. Since colonization speed is bounded by the speed of light, a civilization attaining technological maturity will start with a modest endowment of physical resources (a single planet and perhaps some nearby parts of its solar system), and it will take a very long time—billions of years—before a civilization starting could reach even 1% of its maximum attainable resource base.[21] It is therefore efficiency of use at later times, rather than in the immediate aftermath of the attainment of technological maturity, that matters most for how much value is ultimately realized.

---

[20] It is also one reason why permanent stagnation is an existential risk, although permanent stagnation might also preclude survival beyond the time when the Earth becomes uninhabitable, perhaps around a billion years from now due to increasing solar luminosity (Schroder and Smith 2008).

[21] One potentially significant qualification is that the time to reach the maximum attainable resource base could be shorter if intelligent opposition (such as from extraterrestrial civilizations) emerges that hinders our cosmic expansion.

Furthermore, it might turn out that the ideal way to use most of the cosmic endowment that humanity could eventually secure is to postpone consumption for as long as possible. By conserving our accumulated free energy until the universe is older and colder, we might be able to perform some computations more efficiently.[22] This reinforces the point that it would be a mistake to place too much weight on the amount of value generated shortly after technological maturity when deciding whether some scenario should count as a flawed realization (or a subsequent ruination). It is much more important to get the setup right, in the sense of putting humanity on a track that will eventually garner most of the attainable cosmic resources and put them to near-optimal use. It matters less whether there is a brief delay before that happens—and a delay of even several million years is "brief" in this context (Bostrom 2003).

Even for individual agents, the passage of sidereal time might become less significant after technological maturity. Agents that exist as computational processes in distributed computational hardware have potentially unlimited life spans. The same holds for embodied agents in an era in which physical-repair technologies are sufficiently advanced. The amount of life available to such agents is proportional to the amount of physical resources they control. (A software mind can experience a certain amount of subjective time by running on a slow computer for a long period of sidereal time or, equivalently, by running for a brief period of sidereal time on a fast computer.) Even from a so-called "person-affecting" moral perspective, therefore, when assessing whether a flawed realization has occurred, one should focus not on how much value is created just after the attainment of technological maturity but on whether the conditions created are such as to give a good prospect of realizing a large integral of value over the remainder of the universe's lifetime.

3.2. Some other ethical perspectives

We have thus far considered existential risk from the perspective of utilitarianism (combined with several simplifying assumptions). We may briefly consider how the issue might appear when viewed through the lenses of some other ethical outlooks.

For example, the philosopher Robert Adams outlines a different view on these matters:

[22] There is a minimum entropy cost associated with the erasure of one bit of information, a cost which declines with temperature.

> I believe a better basis for ethical theory in this area can be found in quite a different direction—in a commitment to the future of humanity as a vast project, or network of overlapping projects, that is generally shared by the human race. The aspiration for a better society—more just, more rewarding, and more peaceful—is a part of this project. So are the potentially endless quests for scientific knowledge and philosophical understanding, and the development of artistic and other cultural traditions. This includes the particular cultural traditions to which we belong, in all their accidental historic and ethnic diversity. It also includes our interest in the lives of our children and grandchildren, and the hope that they will be able, in turn, to have the lives of their children and grandchildren as projects. To the extent that a policy or practice seems likely to be favorable or unfavorable to the carrying out of this complex of projects in the nearer or further future, we have reason to pursue or avoid it. … Continuity is as important to our commitment to the project of the future of humanity as it is to our commitment to the projects of our own personal futures. Just as the shape of my whole life, and its connection with my present and past, have an interest that goes beyond that of any isolated experience, so too the shape of human history over an extended period of the future, and its connection with the human present and past, have an interest that goes beyond that of the (total or average) quality of life of a population- at-a-time, considered in isolation from how it got that way.
>
> We owe, I think, some loyalty to this project of the human future. We also owe it a respect that we would owe it even if we were not of the human race ourselves, but beings from another planet who had some understanding of it (Adams 1989, pp. 472-473).

Since an existential catastrophe would either put an end to the project of the future of humanity or drastically curtail its scope for development, we would seem to have a strong *prima facie* reason to avoid it, in Adams' view.

We also note that an existential catastrophe would entail the frustration of many strong preferences, suggesting that from a preference-satisfactionist perspective it would be a bad thing. In a similar vein, an ethical view emphasizing that public policy should be determined through informed democratic deliberation by all stakeholders would favor existential-risk mitigation if we suppose, as is plausible, that a majority of the world's population would come to favor such policies upon reasonable deliberation (even if hypothetical future people are not included as stakeholders). We might also have custodial duties to preserve the inheritance of humanity passed on to us by our ancestors and convey it safely to our descendants.[23] We do not want to be the failing link in the

[23] We might also have responsibilities to nonhuman beings, such as terrestrial (and possible extraterrestrial) animals. Although we are not currently doing much to help them, we have the opportunity to do so in the future. If rendering aid to suffering nonhuman animals in the natural environment is an important value, then

chain of generations, and we ought not to delete or abandon the great epic of human civilization that humankind has been working on for thousands of years, when it is clear that the narrative is far from having reached a natural terminus. Further, many theological perspectives deplore naturalistic existential catastrophes, especially ones induced by human activities: If God created the world and the human species, one would imagine that He might be displeased if we took it upon ourselves to smash His masterpiece (or if, through our negligence or hubris, we allowed it to come to irreparable harm).[24]

We might also consider the issue from a less theoretical standpoint and try to form an evaluation instead by considering analogous cases about which we have definite moral intuitions. Thus, for example, if we feel confident that committing a small genocide is wrong, and that committing a large genocide is no less wrong, we might conjecture that committing omnicide is also wrong.[25] And if we believe we have some moral reason to prevent natural catastrophes that would kill a small number of people, and a stronger moral reason to prevent natural catastrophes that would kill a larger number of people, we might conjecture that we have an even stronger moral reason to prevent catastrophes that would kill the entire human population.

Many different normative perspectives thus concur in their support for existential-risk mitigation, although the degree of badness involved in an existential catastrophe and the priority that existential-risk mitigation should have in our moral economy may vary substantially among different moral theories.[26] Note, however, that it is on no account a *conceptual* truth that existential

---

achieving technological maturity in a manner that fails to produce such aid could count as flawed realization. Cf. McMahan 2010; Pearce 2004.

[24] There could, from a theological perspective, possibly be a special category of existential risks with a different moral status: catastrophes or apocalypses brought about by divine agency, perhaps as just punishment for our sins. A believer might judge such an event as, on balance, good. However, it seems implausible that mere mortals would be able to thwart God if He really wanted to flatten us, so any physical countermeasures we implement against existential risk would presumably be effective only against natural and anthropogenic existential risks, and we might have no reason to hold back on our naturalistic-risk mitigation efforts for fear of frustrating God's designs.

[25] Although omnicide would at least be impartial, by contrast to genocide which is often racist or nationalist.

[26] For example, James Lenman has argued that it is largely a matter of indifference when humankind goes extinct, at least if it does not happen too soon (Lenman 2002).

catastrophes are bad or that reducing existential risk is right. There are possible situations in which the occurrence of one type of existential catastrophe is beneficial—for instance, because it preempts another type of existential catastrophe that would otherwise certainly have occurred and that would have been worse.

3.3. Existential risk and normative uncertainty

Whereas the first two classes of existential risk (human extinction and permanent stagnation) are specified by purely descriptive criteria, the second two (flawed realization and subsequent ruination) are defined normatively. This means that the concept of existential risk is in part an evaluative notion.[27]

Where normative issues are involved, these issues may be contentious. Population ethics, for instance, is fraught with problems about how to deal with various parameters (such as population size, average well-being, thresholds for what counts as a life worth living, inequality, and same vs. different people choices). The evaluation of some scenarios that involve fundamental transformations of human nature is also likely to be contested (Fukuyama 2002; Glover 1984; Kass 2002; Savulescu and Bostrom 2009). Yet not all normative issues are controversial. It will be generally agreed, for example, that a future in which a small human population ekes out a miserable existence within a wrecked ecosystem in the presence of great but unused technological capabilities would count as a dismally flawed realization of humanity's potential and would constitute an existential catastrophe if not reversed.

There will be some types of putative existential risks for which the main uncertainty is normative and others where the main uncertainty is positive. With regard to positive, or descriptive, uncertainty, we saw earlier that if something is not known to be objectively safe, it is risky, at least in the subjective sense relevant to decision making. We can make a parallel move with

[27] In this respect, the concept of existential risk is similar to concepts such as "democracy" and "efficient labor market." A black hole, or a jar of sterile pebbles, is neither a democracy nor an efficient labor market, and we can see that this is so without having to make any normative judgment; yet there may be other objects that cannot be classified as instances or non-instances of these concepts without taking a stand (at least implicitly) on some normative issue.

regard to normative uncertainty. Suppose that some event *X* would reduce biodiversity. Suppose (for the sake of illustration) it is known that *X* would have no other significant consequences and that the reduced biodiversity would not affect humans or any other morally considerable beings. Now, we may be uncertain whether biodiversity has final value (is valuable "for its own sake"). Hence we may be uncertain about whether or not *X* would really be bad. But we can say that if we are not sure whether or not *X* would really be bad (but we *are* sure that *X* would not be good), then *X* is bad in at least the subjective sense relevant to decision making. That is to say, we have reason to prefer that *X* not occur and perhaps reason to take action to prevent *X*.

Exactly *how* one should take into account fundamental moral uncertainty is an open question, but *that* one should do so is clear (Bostrom 2009). We can thus include as existential risks situations in which we know what will happen and we reasonably judge that what will happen *might* be existentially bad—even when there would in fact be nothing bad about the outcome.

We can highlight one consequence of this: Suppose a fully reliable genie offered to grant humanity any wish it might have for its future. Then—even if we could all agree on one such future—we would still face one more potentially serious existential risk: namely, that of choosing unwisely and selecting a future dismally flawed despite appearing, at the moment of our choice, to be the most desirable of all possible futures.

3.4. Keeping our options alive

These reflections on moral uncertainty suggest an alternative, complementary way of looking at existential risk; they also suggest a new way of thinking about the ideal of sustainability. Let me elaborate.

Our present understanding of axiology might well be confused. We may not now know—at least not in concrete detail—what outcomes would count as a big win for humanity; we might not even yet be able to imagine the best ends of our journey. If we are indeed profoundly uncertain about our ultimate aims, then we should recognize that there is a great *option value* in preserving—and ideally improving—our ability to recognize value and to steer the future accordingly. Ensuring that there will be a future version of humanity with great powers and a propensity to use them wisely is plausibly the best way available to us to increase the probability that the future will contain a lot of value. To do this, we must prevent any existential catastrophe.

We thus want to reach a state in which we have (*a*) far greater intelligence, knowledge, and sounder judgment than we currently do; (*b*) far greater ability to solve global-coordination problems; (*c*) far greater technological capabilities and physical resources; and such that (*d*) our values and preferences are not corrupted in the process of getting there (but rather, if possible, improved). Factors *b* and *c* expand the option set available to humanity. Factor *a* increases humanity's ability to predict the outcomes of the available options and understand what each outcome would entail in terms of the realization of human values. Factor *d*, finally, makes humanity more likely to *want* to realize human values.

How we, from our current situation, might best achieve these ends is not obvious (figure 5). While we ultimately need more technology, insight, and coordination, it is not clear that the shortest path to the goal is the best one.

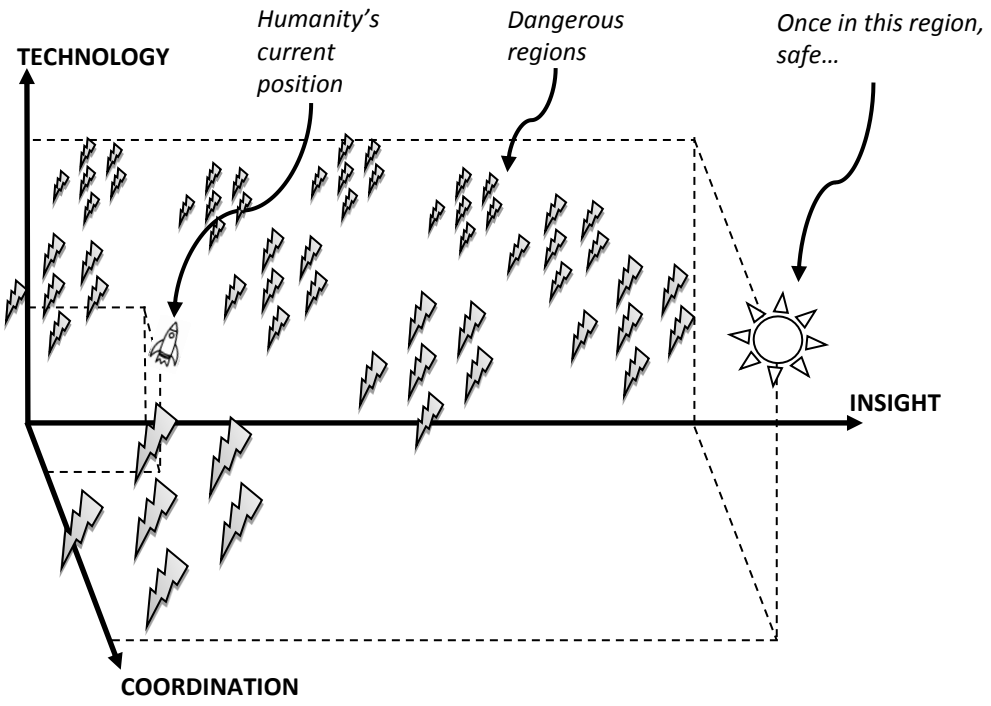

Figure 5: The challenge of finding a safe path. An ideal situation might be one in which we have a very high level of technology, excellent global coordination, and great insight into how our capabilities can be used. It does not follow that getting any amount of additional technology, coordination, or insight is always good for us. Perhaps it is essential that our growth along different dimensions hew to some particular scheme in order for our development to follow a trajectory through the state space that eventually reaches the desired region.

It could turn out, for example, that attaining certain technological capabilities *before* attaining sufficient insight and coordination invariably spells doom for a civilization. One can readily imagine a class of existential-catastrophe scenarios in which some technology is discovered that puts immense destructive power into the hands of a large number of individuals. If there is no effective defense against this destructive power, and no way to prevent individuals from having access to it, then civilization cannot last, since in a sufficiently large population there are bound to be some individuals who will use any destructive power available to them. The discovery of the atomic bomb could have turned out to be like this, except for the fortunate fact that the construction of nuclear weapons requires a special ingredient—weapons-grade fissile material—that is rare and expensive to manufacture. Even so, if we continually sample from the urn of possible technological discoveries before implementing effective means of global coordination, surveillance, and/or restriction of potentially hazardous information, then we risk eventually drawing a black ball: an easy-to-make intervention that causes extremely widespread harm and against which effective defense is infeasible.[28]

We should perhaps therefore not seek *directly* to approximate some state that is "sustainable" in the sense that we could remain in it for some time. Rather, we should focus on getting onto a developmental trajectory that offers a high probability of avoiding existential catastrophe. In other words, our focus should be on maximizing the chances that we will someday attain technological maturity in a way that is not dismally and irremediably flawed. Conditional on that attainment, we have a good chance of realizing our astronomical axiological potential.

To illustrate this point, consider the following analogy. When a rocket stands on the launch pad, it is in a fairly sustainable state. It could remain in its current position for a long time, although it would eventually be destroyed by wind and weather. Another sustainable place for the rocket is in space, where it can travel weightless for a very long time. But when the rocket is in midair, it is in an unsustainable, transitory state: Its engines are blazing and it will soon run out of

[28] Of course, achieving effective global coordination sufficiently strong to continually monitor the entire world population or indefinitely censor any information deemed hazardous by some authority would (at least in the absence of adequate safeguards) create its own very significant existential risks, such as risks of permanent stagnation or flawed realization under some repressive totalitarian regime.

fuel. Returning the rocket to a sustainable state is desirable, but this does not mean that *any* way to render its state more sustainable is desirable. For example, reducing its energy consumption so that it just barely manages to hold stationary might make its state more sustainable in the sense that it can remain in one place for longer; however, when its fuel runs out the rocket will crash to the ground. The best policy for a rocket in midair is, rather, to maintain enough thrust to escape Earth's gravitational field: a strategy that involves entering a *less* sustainable state (consuming fuel faster) in order to later achieve the most desirable sustainable state. That is, instead of seeking to approximate a sustainable *state*, it should pursue a sustainable *trajectory*.

The present human condition is likewise a transitional state. Like the rocket in our analogy, humanity needs to pursue a sustainable trajectory, one that will minimize the risk of existential catastrophe.[29] But unlike the problem of determining the optimum rate of fuel consumption in a rocket, the problem of how to minimize existential risk has no known solution.

4. OUTLOOK

We have seen that reducing existential risk emerges as a dominant priority in many aggregative consequentialist moral theories (and as a very important concern in many other moral theories). The concept of existential risk can thus help the morally or altruistically motivated to identify actions that have the highest expected value. In particular, given certain assumptions, the problem of making the right decision simplifies to that of following the maxipok principle.

[29] Ideally, it would do this while achieving the means to commit collective euthanasia, in the fairly unlikely case that, after long and careful collective deliberation, we should decide that a quick end is preferable to continued existence. That might, however, be a beneficial capability only if we had first attained sufficient wisdom not to exercise it erroneously. We should emphasize the need for continued philosophical deliberation and fostering of conditions that would help us find the truth about central normative issues eventually—as well as the need to avoid irrevocable mistakes in the meantime.

4.1. Barriers to thought and action

In light of this result, which suggests that there may be a very high value in studying existential risks and in analyzing potential mitigation strategies, it is striking how little academic attention these issues have received compared to other topics that are less important (figure 5).[30]

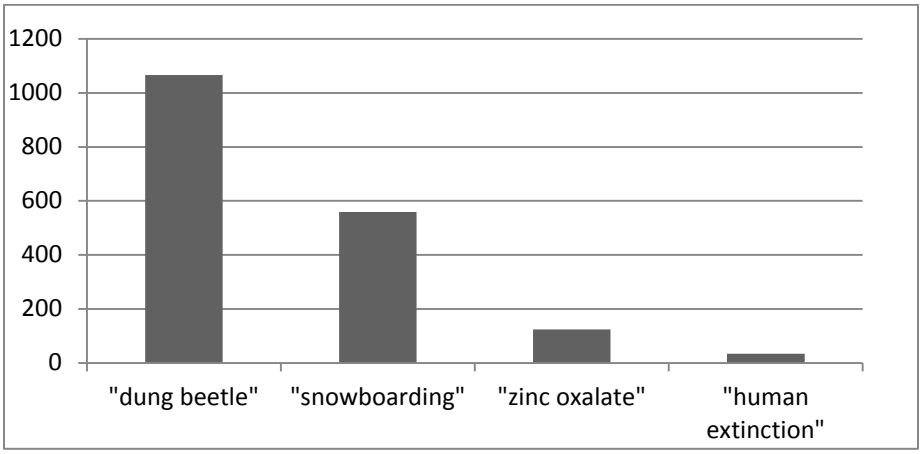

Figure 5: Academic prioritization. Number of academic papers on various topics (listed in Scopus, August 2012).

Many factors conspire against the study and mitigation of existential risks. Research is perhaps inhibited by the multidisciplinary nature of the problem, but also by deeper epistemological issues. The biggest existential risks are not amenable to plug-and-play scientific research methodologies. Furthermore, there are unresolved foundational issues, particularly concerning observation selection theory and population ethics, which are crucial to the assessment of existential

[30] Scholarly treatments of existential risk *per se*, or even of human-extinction risk, are rare (e.g., Bostrom 2002; Leslie 1996; Matheny 2007; Wells 2009). However, a great deal of academic literature bears on individual existential risks or on other specific issues relevant to many existential risks (a few of which are cited throughout this paper). In addition, some recent works take a broad look at global catastrophic risks, though without restricting the focus to existential risks (e.g., Bostrom and Cirkovic 2008; Diamond 2006; Homer-Dixon 2007; Posner 2004; Sunstein 2009; World Economic Forum 2011).

risk; and these theoretical difficulties are compounded by psychological factors that make it difficult to think clearly about issues such as the end of humanity.[31]

If more resources were to be made available to research existential risks, there is a danger that they would flow, with excessive preponderance, to the relatively minor risks that are easier for some established disciplinary community to study using familiar methods, at the expense of far more important risk areas—machine superintelligence, advanced molecular nanotechnology, totalitarianism, risks related to the simulation-hypothesis, or future advances in synthetic biology—which would require a more inconvenient shift in research focus. Another plausible diversion is that research would mainly be directed at global catastrophic risks that involve little or no existential risk.

Mitigation of existential risk is hampered by a lack of understanding, but also by a deficit of motivation. Existential risk mitigation is a global public good (i.e., non-excludable and non-rivalrous), and economic theory suggests that such goods tend to be undersupplied by the market, since each producer of existential safety (even if the producer is a large nation) could capture only a small portion of the value (Feldman 1980; Kaul 1999). In fact, the situation is worse than is the case with many other global public goods in that existential risk reduction is a strongly *transgenerational* (in fact, pan-generational) public good: even a world state may capture only a small fraction of the benefits—those accruing to currently existing people. The quadrillions of happy people who may come to exist in the future if we avoid existential catastrophe would be willing to pay the present generation astronomical sums in return for a slight increase in our efforts to preserve humanity's future, but the mutually beneficial trade is unfortunately prevented by the obvious transaction difficulties.

Moral motivations, too, may fail to measure up to the magnitude of what is at stake. The scope insensitivity of our moral sentiments is likely to be especially pronounced when very large numbers are involved:

---

[31] Relevant issues related to observation selection effects include, among others, the Carter-Leslie doomsday argument, the simulation argument, and "great filter" arguments; see Bostrom 2002, 2003 and 2008; Carter 1983; Cirkovic, Sandberg and Bostrom 2010; Hanson 1998; Leslie 1996; Tegmark and Bostrom 2005. For some relevant issues in moral philosophy, see, e.g., Bostrom 2003 and 2009. For a review of the cognitive-biases literature as it relates to catastrophic risk, see Yudkowsky 2008.

> Substantially larger numbers, such as 500 million deaths, and especially qualitatively different scenarios such as the extinction of the entire human species, seem to trigger a different mode of thinking—enter into a "separate magisterium." People who would never dream of hurting a child hear of an existential risk, and say, "Well, maybe the human species doesn't really deserve to survive." (Yudkowsky 2008, p. 114)

Existential risk requires a proactive approach. The reactive approach—to observe what happens, limit damages, and then implement improved mechanisms to reduce the probability of a repeat occurrence—does not work when there is no opportunity to learn from failure. Instead, we must anticipate emerging dangers, mobilize support for action against hypothetical future harm, and get our precautions sufficiently right the first time. That is a tall order. Few institutions are capable of operating consistently at such a level of effective rationality, and attempts to *imitate* such proactive behavior within less perfect institutions can easily backfire. Speculative risk-mongering could be exploited to rationalize self-serving aggressive action, expansion of costly and potentially oppressive security bureaucracies, or restrictions of civil liberties that keep societies free and sane. The result of false approximations to the rational ideal could easily be a net increase in existential risk.[32]

Multidisciplinary and epistemological challenges, academic distractions and diversions, cognitive biases, free-rider problems, moral lethargy and scope-insensitivity, institutional incompetence, and the political exploitation of unquantifiable threats are thus some of the barriers to effective mitigation. To these we can add the difficulty of achieving required levels of global cooperation. While some existential risks can be tackled unilaterally—any state with a space industry could build a global defense against asteroid impacts—other risks require a joint venture between many states. Management of the global climate may require buy-in by an overwhelming majority of industrialized and industrializing nations. Avoidance of arms races and relinquishment of dangerous directions of technological research may require that *all* states join the effort, since a

[32] A possible way around this problem involves trying to hold the total amount of risk concern roughly constant while allocating a greater proportion of the pot of "fear tokens" or "concern chips" to existential risk. Thus, one might advocate that as we become more concerned about existential risk, we ought simultaneously to become less concerned about smaller risks, such as a few thousand people dying in the odd terrorist attack or natural disaster.

single defector could annul any benefits of collaboration. Some future dangers might even require that each state monitor and regulate every significant group or individual within its territory.[33]

4.2. Grounds for optimism?

A formidable array of obstacles thus clouds the prospect of a clear-headed and effective response to existential risks confronting humanity. Lest the cause be deemed hopeless, we should also take note of some encouraging considerations.

We may note, first, that many of the key concepts and ideas are quite new.[34] Before the conceptual and theoretical foundations were in place, support for efforts to research and mitigate existential risk could not build. In many instances, the underlying scientific, technological, and methodological ideas needed for studying existential risks in a meaningful way have also only recently become available. The delayed start helps explain the still primitive state of the art.

It is arguably only since the detonation of the first atomic bomb in 1945, and the subsequent nuclear buildup during the Cold War, that any significant naturalistic (i.e., non-supernatural) existential risks have arisen—at least if we count only risks over which human beings have some influence.[35] Most of the really big existential risks still seem to lie many years into the future. Until

---

[33] Such internal control within states will become more feasible with advances in surveillance technology. As noted, preventing states with such capabilities from becoming oppressive will present its own set of challenges.

[34] Including the very notion of existential risk (Bostrom 2002).

[35] One could argue that pandemics and close encounters with comets, which occurred repeatedly in human history and elicited strong end-of-the-world forebodings, should count as large early existential risks. Given the limited information then available, it might not have been unreasonable for contemporary observers to assign a significant probability to the end being nigh. Religious doomsday scenarios could also be considered; perhaps it was not unreasonable to believe, on the basis of the then-available evidence, that these risks were real and, moreover, that they could be mitigated through such actions as repentance, prayer, sacrificial offerings, persecution of witches or infidels, and so forth. The first clear-cut scientific existential risk might have arisen with the development of the atomic bomb. Robert Oppenheimer, the scientific leader of the Manhattan Project, ordered a study ahead of the Trinity test to determine whether a nuclear detonation would cause a self-propagating chain of nuclear reactions in Earth's atmosphere. The resulting report may represent the first quantitative risk assessment of human extinction (Manhattan Project 1946).

recently, therefore, there may have been relatively little need to think about existential risk in general and few opportunities for mitigation even if such thinking had taken place.

Public awareness of the global impacts of human activities appears to be increasing. Systems, processes, and risks are studied today from a global perspective by many scholars—environmental scientists, economists, epidemiologists, demographers, and others. Problems such as climate change, cross-border terrorism, and international financial crises direct attention to global interdependency and threats to the global system. The idea of risk in general seems to have risen in prominence.[36] Given these advances in knowledge, methods, and attitudes, the conditions for securing for existential risks the scrutiny they deserve are unprecedentedly propitious.

Opportunities for action may also proliferate. As noted, some mitigation projects can be undertaken unilaterally, and one may expect more such projects as the world becomes richer. Other mitigation projects require wider coordination; in many cases, global coordination. Here, too, some trend lines seem to point to this becoming more feasible over time. There is a long-term historic trend toward increasing scope of political integration—from hunter-gatherer bands to chiefdoms, city states, nation states, and now multinational organizations, regional alliances, various international governance structures, and other aspects of globalization (Wright 1999). Extrapolation of this trend might seem to indicate the eventual creation of a singleton (Bostrom 2006). It is also possible that some of the global movements that emerged over the last half century—in particular the peace movement, the environmentalist movement, and various global justice and human-rights movements—will increasingly take on board more generalized concerns about existential risk.[37]

Furthermore, to the extent that existential-risk mitigation really is a most deserving cause, one may expect that general improvements in society's ability to recognize and act on important truths will differentially funnel resources into existential-risk mitigation. General improvements of

---

[36] Some sociologists have gone so far as to fixate on risk as a central thematic of our age; see, e.g., Beck 1999.

[37] Many peace activists opposing the nuclear arms race during the Cold War explicitly fretted about a nuclear Armageddon that could allegedly end all human life. More recently some environmentalists sounding the alarm about global warming use similarly apocalyptic language. It is unclear, however, to what extent the perceived possibility of a species-ending outcome has been a major motivating force in these cases. Perhaps the amount of concern would be roughly the same even in the face of an iron-clad guarantee that any catastrophe would stop short of human extinction.

this kind might come from many sources, including developments in educational techniques and online collaboration tools, institutional innovations such as prediction markets, advances in science and philosophy, spread of rationality culture, and biological cognitive enhancement.

Finally, it is possible that the cause will at some point receive a boost from the occurrence of a major (non-existential) catastrophe that underscores the precariousness of the present human condition. That would, needless to say, be the worst possible way for our minds to be concentrated—yet one which, in a multidecadal time frame, must be accorded a non-negligible probability of occurrence.[38]

REFERENCES

Adams RM 1989, 'Should ethics be more impersonal? A critical notice of Derek Parfit, *Reasons and persons*', *Philosophical Review*, vol. 98, no. 4, pp. 439-484.

Beck U 1999, *The world risk society*, Polity, Cambridge.

Bostrom N 2002, *Anthropic bias: observation selection effects in science and philosophy*, Routledge, New York.

Bostrom N 2002, 'Existential risks: analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology*, vol. 9, no. 1.

Bostrom N 2003, 'Are you living in a computer simulation?', *Philosophical Quarterly*, vol. 53, no. 211, pp. 243-255.

Bostrom N, 2003, 'Astronomical waste: the opportunity cost of delayed technological development', *Utilitas*, vol 15, no. 3, pp. 308-314.

Bostrom N 2003, 'Infinite ethics', revised version 2009, viewed 15 March 2011, <http://www.nickbostrom.com/ethics/infinite.pdf>.

Bostrom N 2004, 'The future of human evolution', in: Tandy C (ed), *Death and anti-death: two hundred years after Kant, fifty years after Turing*, Ria University Press, Palo Alto, CA, pp. 339-371.

[38] I am grateful for comments and discussion to Seth Baum, Nick Beckstead, Milan Cirkovic, Olle Häggström, Sara Lippincott, Gaverick Matheny, Toby Ord, Derek Parfit, Martin Rees, Rebecca Roache, Anders Sandberg, and Carl Shulman.

Bostrom N 2006, 'What is a singleton?', *Linguistic and Philosophical Investigations*, vol. 5, no. 2, pp. : 48-54.

Bostrom N 2008, 'Where are they? Why I hope the search for extraterrestrial life finds nothing', *MIT Technology Review*, vol. May/June, pp. 72-77.

Bostrom N 2009, 'The future of humanity', in: Olsen J-KB, Selinger E, Riis S (eds), *New Waves in Philosophy of Technology,* Palgrave Macmillan, New York, pp. 186-216.

Bostrom N 2009, 'Moral uncertainty—towards a solution?', in *Overcoming Bias,* viewed 23 March 2011. <http://www.overcomingbias.com/2009/01/moral-uncertainty-towards-a-solution.html>.

Bostrom N 2009, 'Pascal's mugging', *Analysis*, vol. 69, no. 3, pp. 443-445.

Bostrom N, Cirkovic MM (eds) 2008, *Global Catastrophic Risks,* Oxford University Press, Oxford.

Carter B 1983, 'The anthropic principle and its implications for biological evolution', *Philosophical Transactions of the Royal Society*, vol. A 310, pp. 347-363.

Cirkovic MM 2004, 'Forecast for the next eon: applied cosmology and the long-term fate of intelligent beings', *Foundations of Physics*, vol. 34, no. 2, pp. 239-261.

Cirkovic MM, Radujkov M 2001, 'On the maximal quantity of processed information in the physical eschatological context', *Serbian Astronomy Journal*, vol. 163, pp. 53-56.

Cirkovic MM, Sandberg A, Bostrom N 2010, 'Anthropic shadow: observation selection effects and human extinction risks', *Risk Analysis*, vol. 30, no. 10, pp. 1495-1506.

Diamond J 2006, *Collapse: how societies choose to fail or survive,* Penguin, London.

Feldman A 1980, *Welfare economics and social choice theory,* Martinus Nijhoff Publishing, Boston.

Freitas RA 1980, 'A self-reproducing interstellar probe', *Journal of the British Interplanetary Society*, vol. 33, pp. 251-264.

Freitas RA 1999, *Nanomedicine volume I: basic capabilities,* Landes Bioscience, Austin, TX.

Freitas RA 2003, *Nanomedicine volume IIA: biocompatibility*. Landes Bioscience, Austin, TX.

Fukuyama F 2002, *Our posthuman future: consequences of the biotechnology revolution,* Profile, London.

Glover J 1984, *What sort of people should there be?* Penguin, Harmondsworth.

Gott JR, Juric M, Schlegel D, Hoyle F, Vogeley M, Tegmark M, Bahcall N, Brinkmann J 2005, 'A map of the universe', *Astrophysical Journal*, vol. 624, no. 2, pp. 463-483.

Hanson R 1994, 'If uploads come first?', *Extropy*, vol. 6, no. 2, pp. 10-15.

Hanson R 1998, 'Burning the cosmic commons: evolutionary strategies for interstellar colonization',

viewed 3 April 2011, <http://hanson.gmu.edu/filluniv.pdf>.

Hanson R 1998, 'The Great Filter—Are We Almost Past It?', weblog post, 15 September 1998, viewed 24 March 2011, <http://hanson.gmu.edu/greatfilter.html>.

Heyl JS 2005, 'The long-term future of space travel', *Physical Review D,* vol. 72, pp. 1-4.

Homer-Dixon T 2007, *The upside of down: catastrophe, creativity and the renewal of civilization,* Souvenir Press, London.

Kass L 2002, '*Life, liberty and the defense of dignity',* Encounter, San Francisco, CA.

Kaul I 1999, *Global public goods,* Oxford University Press, Oxford.

Krauss LM, Starkman GD 2000, 'Life, the universe, and nothing: life and death in an ever-expanding universe', *Astrophysical Journal*, vol. 531, no. 1, pp. 22-30.

Lenman J 2002, 'On becoming extinct', *Pacific Philosophical Quarterly*, vol. 83, no. 3, pp. 253-269.

Leslie J 1996, *The end of the world: the science and ethics of human extinction,* Routledge, London.

Manhattan Project 1946, *LA-602: ignition of the atmosphere with nuclear bombs*, Library Without Walls Project, Los Alamos National Laboratory, viewed 24 March 2011, <http://www.fas.org/sgp/othergov/doe/lanl/docs1/00329010.pdf>.

Matheny JG, 2007, 'Reducing the risk of human extinction', *Risk Analysis*, vol. 27, no. 5, pp. 1335-1344.

McMahan J 2010, 'The meat eaters', *New York Times*, 19 September 2010, viewed 23 March 2011, <http://opinionator.blogs.nytimes.com/2010/09/19/the-meat-eaters/>.

Moravec H 1988, *Mind children: the future of robot and human intelligence*, Harvard University, Press, Cambridge, MA.

Ord T, Hillerbrand R, Sandberg A 2010, 'Probing the improbable: methodological challenges for risks with low probabilities and high stakes', *Journal of Risk Research*, vol. 13, pp. 191-205.

Parfit D 1984, *Reasons and Persons,* Clarendon Press, Oxford.

Pearce D 2004, 'The hedonistic imperative', viewed 23 March 2011, <http://www.hedweb.com/welcome.htm>.

Posner RA 2004, *Catastrophe*. Oxford University Press, Oxford.

Rawls J 1971, *A theory of justice*. Harvard University Press, Cambridge, MA (revised edition 1999).

Sandberg A, Bostrom N 2008, *Global catastrophic risks survey*, Future of Humanity Institute *Technical*

Report, #2008-1, Oxford, viewed 9 March 2011. <http://www.global-catastrophic-risks.com/docs/2008-1.pdf>.

Sandberg A, Bostrom N 2008, *Whole brain emulation: a roadmap*, Future of Humanity Institute *Technical Report*, #2008-3, Oxford, viewed 22 March 2011, <http://www.fhi.ox.ac.uk/__data/assets/pdf_file/0019/3853/brain-emulation-roadmap-report.pdf>.

Savulescu J, Bostrom N (eds) 2009, *Enhancing Humans,* Oxford University Press, Oxford.

Schroder K-P, Smith R 2008, 'Distant future of the Sun and Earth revisited', *Monthly Notices of the Royal Astronomical Society*, vol. 368, no. 1, pp. 155-163.

Smil, V 2008, *Global catastrophes and trends: the next fifty years,* The MIT Press, Cambridge, MA.

Sunstein C 2009, *Worst-Case scenarios,* Harvard University Press, Cambridge, MA.

Tegmark M, Bostrom N 2005, 'How unlikely is a Doomsday catastrophe?', *Nature*, vol. 438. p. 754.

Tipler F 1980, 'Extraterrestrial intelligent beings do not exist', *Royal Astronomical Society Quarterly Journal*, vol. 21, pp. 267-281.

U.K. Treasury 2006, *Stern review on the economics of climate change*, viewed 9 March 2011, <http://www.hm-treasury.gov.uk/media/8A3/83/Chapter_2_A_-_Technical_Annex.pdf>.

Weitzman, M. L 2009, 'The extreme uncertainty of extreme climate change: an overview and some implications', Harvard University, mimeo, October 2009.

Wells W 2009, *Apocalypse when? Calculating how long the human race will survive,* Praxis, Chichester.

World Economic Forum. *Global risks, 2011*, viewed 24 March 2011, <http://riskreport.weforum.org/>.

Wright R 1999, *Nonzero: the logic of human destiny,* Pantheon Books, New York.

Yudkowsky E 2008, 'Cognitive biases potentially affecting judgment of global risks', in: Bostrom N, Cirkovic MM (eds), *Global catastrophic risks,* Oxford University Press, Oxford, pp. 91-119.