Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards

Nick Bostrom
Faculty of Philosophy, Oxford University

[Reprinted from: *Journal of Evolution and Technology*, Vol. 9, March 2002. First version: 2001]

Abstract

Because of accelerating technological progress, humankind may be rapidly approaching a critical phase in its career. In addition to well-known threats such as nuclear holocaust, the prospects of radically transforming technologies like nanotech systems and machine intelligence present us with unprecedented opportunities and risks. Our future, and whether we will have a future at all, may well be determined by how we deal with these challenges. In the case of radically transforming technologies, a better understanding of the transition dynamics from a human to a "posthuman" society is needed. Of particular importance is to know where the pitfalls are: the ways in which things could go terminally wrong. While we have had long exposure to various personal, local, and endurable global hazards, this paper analyzes a recently emerging category: that of *existential risks*. These are threats that could cause our extinction or destroy the potential of Earth-originating intelligent life. Some of these threats are relatively well known while others, including some of the gravest, have gone almost unrecognized. Existential risks have a cluster of features that make ordinary risk management ineffective. A final section of this paper discusses several ethical and policy implications. A clearer understanding of the threat picture will enable us to formulate better strategies.

1 Introduction

It's dangerous to be alive and risks are everywhere. Luckily, not all risks are equally serious. For present purposes we can use three dimensions to describe the magnitude of a risk: *scope*, *intensity*, and *probability*. By "scope" I mean the size of the group of people that are at risk. By "intensity" I mean how badly each individual in the group would be affected. And by "probability" I mean the best current subjective estimate of the probability of the adverse outcome.[1]

1.1 A typology of risk

We can distinguish six qualitatively distinct types of risks based on their scope and intensity (*figure 1*). The third dimension, probability, can be superimposed on the two

[1] In other contexts, the notion of "best current subjective estimate" could be operationalized as the market betting odds on the corresponding Idea Future's claim [1]. This remark may help to illustrate the intended concept, but it would not serve as a definition. Only a fool would bet on human extinction since there would be no chance of getting paid whether one won or lost.

dimensions plotted in the figure. Other things equal, a risk is more serious if it has a substantial probability and if our actions can make that probability significantly greater or smaller.

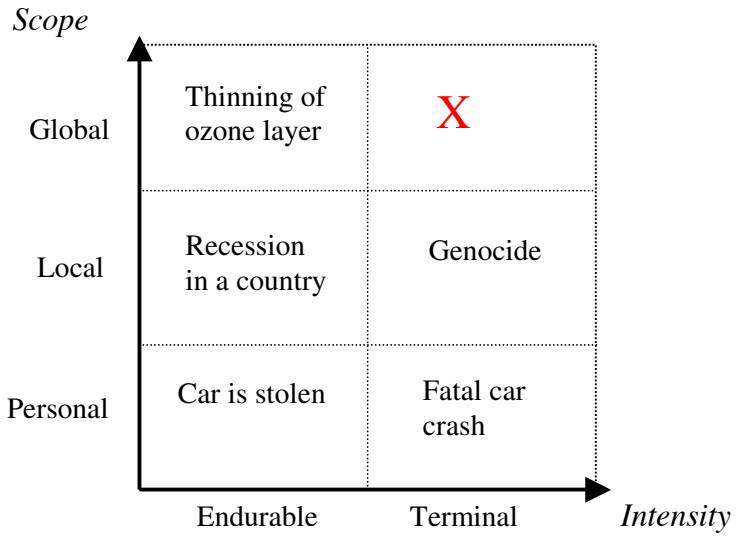

Figure 1. Six risk categories

"Personal", "local", or "global" refer to the size of the population that is directly affected; a global risk is one that affects the whole of humankind (and our successors). "Endurable" vs. "terminal" indicates how intensely the target population would be affected. An endurable risk may cause great destruction, but one can either recover from the damage or find ways of coping with the fallout. In contrast, a terminal risk is one where the targets are either annihilated or irreversibly crippled in ways that radically reduce their potential to live the sort of life they aspire to. In the case of personal risks, for instance, a terminal outcome could for example be death, permanent severe brain injury, or a lifetime prison sentence. An example of a local terminal risk would be genocide leading to the annihilation of a people (this happened to several Indian nations). Permanent enslavement is another example.

1.2 Existential risks

In this paper we shall discuss risks of the sixth category, the one marked with an *X*. This is the category of global, terminal risks. I shall call these *existential risks*.

Existential risks are distinct from global endurable risks. Examples of the latter kind include: threats to the biodiversity of Earth's ecosphere, moderate global warming, global economic recessions (even major ones), and possibly stifling cultural or religious eras such as the "dark ages", even if they encompass the whole global community, provided they are transitory (though see the section on "Shrieks" below). To say that a

particular global risk is endurable is evidently not to say that it is acceptable or not very serious. A world war fought with conventional weapons or a Nazi-style *Reich* lasting for a decade would be extremely horrible events even though they would fall under the rubric of endurable global risks since humanity could eventually recover. (On the other hand, they could be a *local* terminal risk for many individuals and for persecuted ethnic groups.)

I shall use the following definition of existential risks:

> *Existential risk* – One where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential.

An existential risk is one where humankind as a whole is imperiled. Existential disasters have major adverse consequences for the course of human civilization for all time to come.

2 The unique challenge of existential risks

Risks in this sixth category are a recent phenomenon. This is part of the reason why it is useful to distinguish them from other risks. We have not evolved mechanisms, either biologically or culturally, for managing such risks. Our intuitions and coping strategies have been shaped by our long experience with risks such as dangerous animals, hostile individuals or tribes, poisonous foods, automobile accidents, Chernobyl, Bhopal, volcano eruptions, earthquakes, draughts, World War I, World War II, epidemics of influenza, smallpox, black plague, and AIDS. These types of disasters have occurred many times and our cultural attitudes towards risk have been shaped by trial-and-error in managing such hazards. But tragic as such events are to the people immediately affected, in the big picture of things – from the perspective of humankind as a whole – even the worst of these catastrophes are mere ripples on the surface of the great sea of life. They haven't significantly affected the total amount of human suffering or happiness or determined the long-term fate of our species.

With the exception of a species-destroying comet or asteroid impact (an extremely rare occurrence), there were probably no significant existential risks in human history until the mid-twentieth century, and certainly none that it was within our power to do something about.

The first manmade existential risk was the inaugural detonation of an atomic bomb. At the time, there was some concern that the explosion might start a runaway chain-reaction by "igniting" the atmosphere. Although we now know that such an outcome was physically impossible, it qualifies as an existential risk that was present at the time. For there to be a risk, given the knowledge and understanding available, it

suffices that there is some *subjective probability* of an adverse outcome, even if it later turns out that objectively there was no chance of something bad happening. *If we don't know whether something is objectively risky or not, then it is risky in the subjective sense*. The subjective sense is of course what we must base our decisions on.[2] At any given time we must use *our best current subjective estimate* of what the objective risk factors are.[3]

A much greater existential risk emerged with the build-up of nuclear arsenals in the US and the USSR. An all-out nuclear war was a possibility with both a substantial probability and with consequences that *might* have been persistent enough to qualify as global and terminal. There was a real worry among those best acquainted with the information available at the time that a nuclear Armageddon would occur and that it might annihilate our species or permanently destroy human civilization.[4] Russia and the US retain large nuclear arsenals that could be used in a future confrontation, either accidentally or deliberately. There is also a risk that other states may one day build up large nuclear arsenals. Note however that a smaller nuclear exchange, between India and Pakistan for instance, is not an existential risk, since it would not destroy or thwart humankind's potential permanently. Such a war might however be a local terminal risk for the cities most likely to be targeted. Unfortunately, we shall see that nuclear Armageddon and comet or asteroid strikes are mere preludes to the existential risks that we will encounter in the 21st century.

The special nature of the challenges posed by existential risks is illustrated by the following points:

- Our approach to existential risks cannot be one of trial-and-error. There is no opportunity to learn from errors. The reactive approach – see what happens, limit damages, and learn from experience – is unworkable. Rather, we must take a proactive approach. This requires *foresight* to anticipate new types of threats and a willingness to take decisive *preventive action* and to bear the costs (moral and economic) of such actions.

[2] This can be seen as the core wisdom of the so-called Precautionary Principle [2]. Any stronger interpretation of the principle, for instance in terms of where the burden of proof lies in disputes about introducing a risky new procedure, can easily become unreasonably simplistic [3].

[3] On the distinction between objective and subjective probability, see e.g. [4-6]. For a classic treatment of decision theory, see [7].

[4] President Kennedy is said to have at one point estimated the probability of a nuclear war between the US and the USSR to be "somewhere between one out of three and even" ([8], p. 110; see also [9], ch. 2). John von Neumann (1903-1957), the eminent mathematician and one of the founders of game theory and computer science and who as chairman of the Air Force Strategic Missiles Evaluation Committee was a key architect of early US nuclear strategy, is reported to have said it was "absolutely certain (1) that there would be a nuclear war; and (2) that everyone would die in it" [10], p. 114.

- We cannot necessarily rely on the institutions, moral norms, social attitudes or national security policies that developed from our experience with managing other sorts of risks. Existential risks are a different kind of beast. We might find it hard to take them as seriously as we should simply because we have never yet witnessed such disasters.[5] Our collective fear-response is likely ill calibrated to the magnitude of threat.

- Reductions in existential risks are *global public goods* [13] and may therefore be undersupplied by the market [14]. Existential risks are a menace for everybody and may require acting on the international plane. Respect for national sovereignty is not a legitimate excuse for failing to take countermeasures against a major existential risk.

- If we take into account the welfare of future generations, the harm done by existential risks is multiplied by another factor, the size of which depends on whether and how much we discount future benefits [15,16].

In view of its undeniable importance, it is surprising how little systematic work has been done in this area. Part of the explanation may be that many of the gravest risks stem (as we shall see) from anticipated future technologies that we have only recently begun to understand. Another part of the explanation may be the unavoidably interdisciplinary and speculative nature of the subject. And in part the neglect may also be attributable to an aversion against thinking seriously about a depressing topic. The point, however, is not to wallow in gloom and doom but simply to take a sober look at what could go wrong so we can create responsible strategies for improving our chances of survival. In order to do that, we need to know where to focus our efforts.

3 Classification of existential risks

We shall use the following four categories to classify existential risks[6]:

[5] As it applies to the human species, that is. Extinction of other species is commonplace. It is estimated that 99% of all species that ever lived on Earth are extinct. We can also gain some imaginative acquaintance with existential disasters through works of fiction. Although there seems to be a bias towards happy endings, there are exceptions such as the film *Dr. Strangelove* [11] and Nevil Shute's poignant novel *On the Beach* [12]. Moreover, in the case of some existential risks (e.g. species-destroying meteor impact), we do have experience of milder versions thereof (e.g. impacts by smaller meteors) that helps us quantify the probability of the larger event. But for most of the serious existential risks, there is no precedent.

[6] The terminology is inspired by the famous lines of T. S. Eliot:

> This is the way the world ends
> Not with a bang but a whimper

Bangs – Earth-originating intelligent life goes extinct in relatively sudden disaster resulting from either an accident or a deliberate act of destruction.

Crunches – The potential of humankind to develop into posthumanity[7] is permanently thwarted although human life continues in some form.

Shrieks – Some form of posthumanity is attained but it is an extremely narrow band of what is possible and desirable.

Whimpers – A posthuman civilization arises but evolves in a direction that leads gradually but irrevocably to either the complete disappearance of the things we value or to a state where those things are realized to only a minuscule degree of what could have been achieved.

Armed with this taxonomy, we can begin to analyze the most likely scenarios in each category. The definitions will also be clarified as we proceed.

4 Bangs

This is the most obvious kind of existential risk. It is conceptually easy to understand. Below are some possible ways for the world to end in a bang.[8] I have tried to rank them roughly in order of how probable they are, in my estimation, to cause the extinction of Earth-originating intelligent life; but my intention with the ordering is more to provide a basis for further discussion than to make any firm assertions.

4.1 Deliberate misuse of nanotechnology

In a mature form, molecular nanotechnology will enable the construction of bacterium-scale self-replicating mechanical robots that can feed on dirt or other organic matter [22-25]. Such replicators could eat up the biosphere or destroy it by other means such as by poisoning it, burning it, or blocking out sunlight. A person of malicious intent in

---

(From "The Hollow Men")

and also by the title of philosopher John Earman's book on the general theory of relativity [17]. For some general desiderata in classifying risks, see [18].

[7] The words "Posthumanity" and "posthuman civilization" are used to denote a society of technologically highly enhanced beings (with much greater intellectual and physical capacities, much longer life-spans, etc.) that we might one day be able to become [19].

[8] Some of these are discussed in more detail in the first two chapters of John Leslie's excellent book [9]; some are briefly discussed in [20]. The recent controversy around Bill Joy's article in *Wired* [21] also drew attention to some of these issues.

possession of this technology might cause the extinction of intelligent life on Earth by releasing such nanobots into the environment.[9]

The technology to produce a destructive nanobot seems considerably easier to develop than the technology to create an effective defense against such an attack (a global nanotech immune system, an "active shield" [23]). It is therefore likely that there will be a period of vulnerability during which this technology must be prevented from coming into the wrong hands. Yet the technology could prove hard to regulate, since it doesn't require rare radioactive isotopes or large, easily identifiable manufacturing plants, as does production of nuclear weapons [23].

Even if effective defenses against a limited nanotech attack are developed before dangerous replicators are designed and acquired by suicidal regimes or terrorists, there will still be the danger of an arms race between states possessing nanotechnology. It has been argued [26] that molecular manufacturing would lead to both arms race instability and crisis instability, to a higher degree than was the case with nuclear weapons. Arms race instability means that there would be dominant incentives for each competitor to escalate its armaments, leading to a runaway arms race. Crisis instability means that there would be dominant incentives for striking first. Two roughly balanced rivals acquiring nanotechnology would, on this view, begin a massive buildup of armaments and weapons development programs that would continue until a crisis occurs and war breaks out, potentially causing global terminal destruction. That the arms race could have been predicted is no guarantee that an international security system will be created ahead of time to prevent this disaster from happening. The nuclear arms race between the US and the USSR was predicted but occurred nevertheless.

4.2 Nuclear holocaust

The US and Russia still have huge stockpiles of nuclear weapons. But would an all-out nuclear war really exterminate humankind? Note that: (i) For there to be an existential risk it suffices that we can't be sure that it wouldn't. (ii) The climatic effects of a large nuclear war are not well known (there is the possibility of a nuclear winter). (iii) Future arms races between other nations cannot be ruled out and these could lead to even greater arsenals than those present at the height of the Cold War. The world's supply of plutonium has been increasing steadily to about two thousand tons, some ten times as much as remains tied up in warheads ([9], p. 26). (iv) Even if some humans survive the short-term effects of a nuclear war, it could lead to the collapse of civilization. A human race living under stone-age conditions may or may not be more resilient to extinction than other animal species.

[9] Nanotechnology, of course, also holds huge potential for benefiting medicine, the environment, and the economy in general, but that is not the side of the coin that we are studying here.

4.3 We're living in a simulation and it gets shut down

A case can be made that the hypothesis that we are living in a computer simulation should be given a significant probability [27]. The basic idea behind this so-called "Simulation argument" is that vast amounts of computing power may become available in the future (see e.g. [28,29]), and that it could be used, among other things, to run large numbers of fine-grained simulations of past human civilizations. Under some not-too-implausible assumptions, the result can be that almost all minds like ours are simulated minds, and that we should therefore assign a significant probability to being such computer-emulated minds rather than the (subjectively indistinguishable) minds of originally evolved creatures. And if we are, we suffer the risk that the simulation may be shut down at any time. A decision to terminate our simulation may be prompted by our actions or by exogenous factors.

While to some it may seem frivolous to list such a radical or "philosophical" hypothesis next the concrete threat of nuclear holocaust, we must seek to base these evaluations on reasons rather than untutored intuition. Until a refutation appears of the argument presented in [27], it would intellectually dishonest to neglect to mention simulation-shutdown as a potential extinction mode.

4.4 Badly programmed superintelligence

When we create the first superintelligent entity [28-34], we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so. For example, we could mistakenly elevate a subgoal to the status of a supergoal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device, in the process killing the person who asked the question. (For further analysis of this, see [35].)

4.5 Genetically engineered biological agent

With the fabulous advances in genetic technology currently taking place, it may become possible for a tyrant, terrorist, or lunatic to create a doomsday virus, an organism that combines long latency with high virulence and mortality [36].

Dangerous viruses can even be spawned unintentionally, as Australian researchers recently demonstrated when they created a modified mousepox virus with 100% mortality while trying to design a contraceptive virus for mice for use in pest control [37]. While this particular virus doesn't affect humans, it is suspected that an analogous alteration would increase the mortality of the human smallpox virus. What underscores the future hazard here is that the research was quickly published in the open scientific literature [38]. It is hard to see how information generated in open biotech research

programs could be contained no matter how grave the potential danger that it poses; and the same holds for research in nanotechnology.

Genetic medicine will also lead to better cures and vaccines, but there is no guarantee that defense will always keep pace with offense. (Even the accidentally created mousepox virus had a 50% mortality rate on vaccinated mice.) Eventually, worry about biological weapons may be put to rest through the development of nanomedicine, but while nanotechnology has enormous long-term potential for medicine [39] it carries its own hazards.

4.6 Accidental misuse of nanotechnology ("gray goo")

The possibility of accidents can never be completely ruled out. However, there are many ways of making sure, through responsible engineering practices, that species-destroying accidents do not occur. One could avoid using self-replication; one could make nanobots dependent on some rare feedstock chemical that doesn't exist in the wild; one could confine them to sealed environments; one could design them in such a way that any mutation was overwhelmingly likely to cause a nanobot to completely cease to function [40]. Accidental misuse is therefore a smaller concern than malicious misuse [23,25,41].

However, the distinction between the accidental and the deliberate can become blurred. While "in principle" it seems possible to make terminal nanotechnological accidents extremely improbable, the actual circumstances may not permit this ideal level of security to be realized. Compare nanotechnology with nuclear technology. From an engineering perspective, it is of course perfectly possible to use nuclear technology only for peaceful purposes such as nuclear reactors, which have a zero chance of destroying the whole planet. Yet in practice it may be very hard to avoid nuclear technology also being used to build nuclear weapons, leading to an arms race. With large nuclear arsenals on hair-trigger alert, there is inevitably a significant risk of accidental war. The same can happen with nanotechnology: it may be pressed into serving military objectives in a way that carries unavoidable risks of serious accidents.

In some situations it can even be strategically advantageous to *deliberately* make one's technology or control systems risky, for example in order to make a "threat that leaves something to chance" [42].

4.7 Something unforeseen

We need a catch-all category. It would be foolish to be confident that we have already imagined and anticipated all significant risks. Future technological or scientific developments may very well reveal novel ways of destroying the world.

Some *foreseen* hazards (hence not members of the current category) which have been excluded from the list of bangs on grounds that they seem too unlikely to cause a global terminal disaster are: solar flares, supernovae, black hole explosions or mergers,

gamma-ray bursts, galactic center outbursts, supervolcanos, loss of biodiversity, buildup of air pollution, gradual loss of human fertility, and various religious doomsday scenarios. The hypothesis that we will one day become "illuminated" and commit collective suicide or stop reproducing, as supporters of VHEMT (The Voluntary Human Extinction Movement) hope [43], appears unlikely. If it really were better not to exist (as Silenus told king Midas in the Greek myth, and as Arthur Schopenhauer argued [44] although for reasons specific to his philosophical system he didn't advocate suicide), then we should not count this scenario as an existential disaster. The assumption that it is not worse to be alive should be regarded as an implicit assumption in the definition of Bangs. *Erroneous* collective suicide is an existential risk albeit one whose probability seems extremely slight. (For more on the ethics of human extinction, see chapter 4 of [9].)

4.8 Physics disasters

The Manhattan Project bomb-builders' concern about an A-bomb-derived atmospheric conflagration has contemporary analogues.

There have been speculations that future high-energy particle accelerator experiments may cause a breakdown of a metastable vacuum state that our part of the cosmos might be in, converting it into a "true" vacuum of lower energy density [45]. This would result in an expanding bubble of total destruction that would sweep through the galaxy and beyond at the speed of light, tearing all matter apart as it proceeds.

Another conceivability is that accelerator experiments might produce negatively charged stable "strangelets" (a hypothetical form of nuclear matter) or create a mini black hole that would sink to the center of the Earth and start accreting the rest of the planet [46].

These outcomes *seem* to be impossible given our best current physical theories. But the reason we do the experiments is precisely that we don't really know what will happen. A more reassuring argument is that the energy densities attained in present day accelerators are far lower than those that occur naturally in collisions between cosmic rays [46,47]. It's possible, however, that factors other than energy density are relevant for these hypothetical processes, and that those factors will be brought together in novel ways in future experiments.

The main reason for concern in the "physics disasters" category is the meta-level observation that discoveries of all sorts of weird physical phenomena are made all the time, so even if right now all the particular physics disasters we have conceived of were absurdly improbable or impossible, there could be other more realistic failure-modes waiting to be uncovered. The ones listed here are merely illustrations of the general case.

4.9 Naturally occurring disease

What if AIDS was as contagious as the common cold?

There are several features of today's world that may make a global pandemic more likely than ever before. Travel, food-trade, and urban dwelling have all increased dramatically in modern times, making it easier for a new disease to quickly infect a large fraction of the world's population.

4.10 Asteroid or comet impact

There is a real but very small risk that we will be wiped out by the impact of an asteroid or comet [48].

In order to cause the extinction of human life, the impacting body would probably have to be greater than 1 km in diameter (and probably 3 - 10 km). There have been at least five and maybe well over a dozen mass extinctions on Earth, and at least some of these were probably caused by impacts ([9], pp. 81f.). In particular, the K/T extinction 65 million years ago, in which the dinosaurs went extinct, has been linked to the impact of an asteroid between 10 and 15 km in diameter on the Yucatan peninsula. It is estimated that a 1 km or greater body collides with Earth about once every 0.5 million years.[10] We have only catalogued a small fraction of the potentially hazardous bodies.

If we were to detect an approaching body in time, we would have a good chance of diverting it by intercepting it with a rocket loaded with a nuclear bomb [49].

4.11 Runaway global warming

One scenario is that the release of greenhouse gases into the atmosphere turns out to be a strongly self-reinforcing feedback process. Maybe this is what happened on Venus, which now has an atmosphere dense with CO_2 and a temperature of about 450^{O} C. Hopefully, however, we will have technological means of counteracting such a trend by the time it would start getting truly dangerous.

5 Crunches

While some of the events described in the previous section would be certain to actually wipe out Homo sapiens (e.g. a breakdown of a meta-stable vacuum state) others could potentially be survived (such as an all-out nuclear war). If modern civilization were to collapse, however, it is not completely certain that it would arise again even if the human species survived. We *may* have used up too many of the easily available resources a primitive society would need to use to work itself up to our level of technology. A primitive human society may or may not be more likely to face extinction than any other animal species. But let's not try that experiment.

[10] By comparison, the Tunguska event in 1908 was caused by a body about 60 meters in diameter, producing a yield of 2 megatons TNT (the Hiroshima bomb had a yield of 2 kilotons) and felling trees within a 40 km radius.

If the primitive society lives on but fails to ever get back to current technological levels, let alone go beyond it, then we have an example of a crunch. Here are some potential causes of a crunch:

5.1 Resource depletion or ecological destruction

The natural resources needed to sustain a high-tech civilization are being used up. If some other cataclysm destroys the technology we have, it may not be possible to climb back up to present levels if natural conditions are less favorable than they were for our ancestors, for example if the most easily exploitable coal, oil, and mineral resources have been depleted. (On the other hand, if plenty of information about our technological feats is preserved, that could make a rebirth of civilization easier.)

5.2 Misguided world government or another static social equilibrium stops technological progress

One could imagine a fundamentalist religious or ecological movement one day coming to dominate the world. If by that time there are means of making such a world government stable against insurrections (by advanced surveillance or mind-control technologies), this might permanently put a lid on humanity's potential to develop to a posthuman level. Aldous Huxley's *Brave New World* is a well-known scenario of this type [50].

A world government may not be the only form of stable social equilibrium that could permanently thwart progress. Many regions of the world today have great difficulty building institutions that can support high growth. And historically, there are many places where progress stood still or retreated for significant periods of time. Economic and technological progress may not be as inevitable as is appears to us.

5.3 "Dysgenic" pressures

It is possible that advanced civilized society is dependent on there being a sufficiently large fraction of intellectually talented individuals. Currently it seems that there is a negative correlation in some places between intellectual achievement and fertility. If such selection were to operate over a long period of time, we might evolve into a less brainy but more fertile species, *homo philoprogenitus* ("lover of many offspring").

However, contrary to what such considerations might lead one to suspect, IQ scores have actually been increasing dramatically over the past century. This is known as the Flynn effect; see e.g. [51,52]. It's not yet settled whether this corresponds to real gains in important intellectual functions.

Moreover, genetic engineering is rapidly approaching the point where it will become possible to give parents the choice of endowing their offspring with genes that correlate with intellectual capacity, physical health, longevity, and other desirable traits.

In any case, the time-scale for human natural genetic evolution seems much too grand for such developments to have any significant effect before other developments will have made the issue moot [19,39].

5.4 Technological arrest

The sheer technological difficulties in making the transition to the posthuman world might turn out to be so great that we never get there.

5.5 Something unforeseen[11]

As before, a catch-all.

Overall, the probability of a crunch seems much smaller than that of a bang. We should keep the possibility in mind but not let it play a dominant role in our thinking at this point. If technological and economical development were to slow down substantially for some reason, then we would have to take a closer look at the crunch scenarios.

6 Shrieks

Determining which scenarios are shrieks is made more difficult by the inclusion of the notion of *desirability* in the definition. Unless we know what is "desirable", we cannot tell which scenarios are shrieks. However, there are some scenarios that would count as shrieks under most reasonable interpretations.

6.1 Take-over by a transcending upload

Suppose uploads come before human-level artificial intelligence. An upload is a mind that has been transferred from a biological brain to a computer that emulates the computational processes that took place in the original biological neural network [19,33,53,54]. A successful uploading process would preserve the original mind's memories, skills, values, and consciousness. Uploading a mind will make it much easier to enhance its intelligence, by running it faster, adding additional computational resources, or streamlining its architecture. One could imagine that enhancing an upload beyond a certain point will result in a positive feedback loop, where the enhanced upload is able to figure out ways of making itself even smarter; and the smarter successor version is in turn even better at designing an improved version of itself, and so on. If this runaway process is sudden, it could result in one upload reaching superhuman levels of intelligence while everybody else remains at a roughly human level. Such enormous

[11] It is questionable whether a badly programmed superintelligence that decided to hold humanity back indefinitely could count as a whimper. The superintelligence would have to be of such a limited nature that it wouldn't itself count as some form of posthumanity; otherwise this would be a shriek.

intellectual superiority may well give it correspondingly great power. It could rapidly invent new technologies or perfect nanotechnological designs, for example. If the transcending upload is bent on preventing others from getting the opportunity to upload, it might do so.

The posthuman world may then be a reflection of one particular egoistical upload's preferences (which in a worst case scenario would be worse than worthless). Such a world may well be a realization of only a tiny part of what would have been possible and desirable. This end is a shriek.

6.2 Flawed superintelligence

Again, there is the possibility that a badly programmed superintelligence takes over and implements the faulty goals it has erroneously been given.

6.3 Repressive totalitarian global regime

Similarly, one can imagine that an intolerant world government, based perhaps on mistaken religious or ethical convictions, is formed, is stable, and decides to realize only a very small part of all the good things a posthuman world could contain.

Such a world government could conceivably be formed by a small group of people if they were in control of the first superintelligence and could select its goals. If the superintelligence arises suddenly and becomes powerful enough to take over the world, the posthuman world may reflect only the idiosyncratic values of the owners or designers of this superintelligence. Depending on what those values are, this scenario would count as a shriek.

6.4 Something unforeseen.[12]

The catch-all.

These shriek scenarios appear to have substantial probability and thus should be taken seriously in our strategic planning.

One could argue that one value that makes up a large portion of what we would consider desirable in a posthuman world is that it contains as many as possible of those persons who are currently alive. After all, many of us want very much not to die (at least not yet) and to have the chance of becoming posthumans. If we accept this, then *any* scenario in which the transition to the posthuman world is delayed for long enough that almost all current humans are dead before it happens (assuming they have not been successfully preserved via cryonics arrangements [53,57]) would be a shriek. Failing a

[12] I regard the hypothesis (common in the mass media and defended e.g. in [55]; see also [56]) that we will be exterminated in a conventional war between the human species and a population of roughly human-equivalent human-made robots as extremely small.

breakthrough in life-extension or widespread adoption of cryonics, then even a smooth transition to a fully developed posthuman eighty years from now would constitute a major existential risk, *if* we define "desirable" with special reference to the people who are currently alive. This "if", however, is loaded with a profound axiological problem that we shall not try to resolve here.

7 Whimpers

If things go well, we may one day run up against fundamental physical limits. Even though the universe appears to be infinite [58,59], the portion of the universe that we could potentially colonize is (given our admittedly very limited current understanding of the situation) finite [60], and we will therefore eventually exhaust all available resources or the resources will spontaneously decay through the gradual decrease of negentropy and the associated decay of matter into radiation. But here we are talking astronomical time-scales. An ending of this sort may indeed be the best we can hope for, so it would be misleading to count it as an existential risk. It does not qualify as a whimper because humanity could on this scenario have realized a good part of its potential.

Two whimpers (apart form the usual catch-all hypothesis) appear to have significant probability:

7.1 Our potential or even our core values are eroded by evolutionary development

This scenario is conceptually more complicated than the other existential risks we have considered (together perhaps with the "We are living in a simulation that gets shut down" bang scenario). It is explored in more detail in a companion paper [61]. An outline of that paper is provided in an Appendix.

A related scenario is described in [62], which argues that our "cosmic commons" could be burnt up in a colonization race. Selection would favor those replicators that spend *all* their resources on sending out further colonization probes [63].

Although the time it would take for a whimper of this kind to play itself out may be relatively long, it could still have important policy implications because near-term choices may determine whether we will go down a track [64] that inevitably leads to this outcome. Once the evolutionary process is set in motion or a cosmic colonization race begun, it could prove difficult or impossible to halt it [65]. It may well be that the only feasible way of avoiding a whimper is to prevent these chains of events from ever starting to unwind.

7.2 Killed by an extraterrestrial civilization

The probability of running into aliens any time soon appears to be very small (see section on evaluating probabilities below, and also [66,67]).

If things go well, however, and we develop into an intergalactic civilization, we may one day in the distant future encounter aliens. If they were hostile and if (for some unknown reason) they had significantly better technology than we will have by then, they may begin the process of conquering us. Alternatively, if they trigger a phase transition of the vacuum through their high-energy physics experiments (see the Bangs section) we may one day face the consequences. Because the spatial extent of our civilization at that stage would likely be very large, the conquest or destruction would take relatively long to complete, making this scenario a whimper rather than a bang.

7.3 Something unforeseen

The catch-all hypothesis.

The first of these whimper scenarios should be a weighty concern when formulating long-term strategy. Dealing with the second whimper is something we can safely delegate to future generations (since there's nothing we can do about it now anyway).

8 Assessing the probability of existential risks

8.1 Direct versus indirect methods

There are two complementary ways of estimating our chances of creating a posthuman world. What we could call the *direct way* is to analyze the various specific failure-modes, assign them probabilities, and then subtract the sum of these disaster-probabilities from one to get the success-probability. In doing so, we would benefit from a detailed understanding of how the underlying causal factors will play out. For example, we would like to know the answers to questions such as: How much harder is it to design a foolproof global nanotech immune system than it is to design a nanobot that can survive and reproduce in the natural environment? How feasible is it to keep nanotechnology strictly regulated for a lengthy period of time (so that nobody with malicious intentions gets their hands on an assembler that is not contained in a tamperproof sealed assembler lab [23])? How likely is it that superintelligence will come before advanced nanotechnology? We can make guesses about these and other relevant parameters and form an estimate that basis; and we can do the same for the other existential risks that we have outlined above. (I have tried to indicate the approximate relative probability of the various risks in the rankings given in the previous four sections.)

Secondly, there is the *indirect way*. There are theoretical constraints that can be brought to bear on the issue, based on some general features of the world in which we live. There is only small number of these, but they are important because they do not rely on making a lot of guesses about the details of future technological and social developments:

8.2 The Fermi Paradox

The Fermi Paradox refers to the question mark that hovers over the data point that we have seen no signs of extraterrestrial life [68]. This tells us that it is not the case that life evolves on a significant fraction of Earth-like planets and proceeds to develop advanced technology, using it to colonize the universe in ways that would have been detected with our current instrumentation. There must be (at least) one Great Filter – an evolutionary step that is extremely improbable – somewhere on the line between Earth-like planet and colonizing-in-detectable-ways civilization [69]. If the Great Filter isn't in our past, we must fear it in our (near) future. Maybe almost every civilization that develops a certain level of technology causes its own extinction.

Luckily, what we know about our evolutionary past is consistent with the hypothesis that the Great Filter is behind us. There are several plausible candidates for evolutionary steps that may be sufficiently improbable to explain why we haven't seen or met any extraterrestrials, including the emergence of the first organic self-replicators, the transition from prokaryotes to eukaryotes, to oxygen breathing, to sexual reproduction, and possibly others.[13] The upshot is that with our currant knowledge in evolutionary biology, Great Filter arguments cannot tell us very much about how likely we are to become posthuman, although they may give us subtle hints [66,70-72].

This would change dramatically if we discovered traces of independently evolved life (whether extinct or not) on other planets. Such a discovery would be bad news. Finding a relatively advanced life-form (multicellular organisms) would be especially depressing.

[13] These are plausible candidates for difficult, critical steps (perhaps requiring simultaneous multi-loci mutations or other rare coincidences) primarily because they took a very long time (by contrast, for instance, of the evolution of Homo sapiens sapiens from our humanoid ancestors). Yet the duration of a step is not always good reason for thinking the step improbable. For example, oxygen breathing took a long time to evolve, but this is not a ground for thinking that it was a difficult step. Oxygen breathing became adaptive only after there were significant levels of free oxygen in the atmosphere, and it took anaerobic organisms hundreds of millions of years to produce enough oxygen to satiate various oxygen sinks and raise the levels of atmospheric oxygen to the required levels. This process was very slow but virtually guaranteed to run to completion eventually, so it would be a mistake to infer that the evolution of oxygen breathing and the concomitant Cambrian explosion represent a hugely difficult step in human evolution.

8.3 Observation selection effects

The theory of observation selection effects may tell us what we should expect to observe given some hypothesis about the distribution of observers in the world. By comparing such predictions to our actual observations, we get probabilistic evidence for or against various hypotheses.

One attempt to apply such reasoning to predicting our future prospects is the so-called Doomsday argument [9,73].[14] It purports to show that we have systematically underestimated the probability that humankind will go extinct relatively soon. The idea, in its simplest form, is that we should think of ourselves as in some sense random samples from the set of all observers in our reference class, and we would be more likely to live as early as we do if there were not a very great number of observers in our reference class living later than us. The Doomsday argument is highly controversial, and I have argued elsewhere that although it may be theoretically sound, some of its applicability conditions are in fact not satisfied, so that applying it to our actual case would be a mistake [75,76].

Other anthropic arguments may be more successful: the argument based on the Fermi-paradox is one example and the next section provides another. In general, one lesson is that we should be careful not to use the fact that life on Earth has survived up to this day and that our humanoid ancestors didn't go extinct in some sudden disaster to infer that that Earth-bound life and humanoid ancestors are highly resilient. Even if on the vast majority of Earth-like planets life goes extinct before intelligent life forms evolve, we should still expect to find ourselves on one of the exceptional planets that were lucky enough to escape devastation.[15] In this case, our past success provides no ground for expecting success in the future.

The field of observation selection effects is methodologically very complex [76,78,79] and more foundational work is needed before we can be confident that we really understand how to reason about these things. There may well be further lessons from this domain that we haven't yet been clever enough to comprehend.

8.4 The Simulation argument

Most people don't believe that they are currently living in a computer simulation. I've recently shown (using only some fairly uncontroversial parts of the theory of observation selection effects) that this commits one to the belief that either we are almost certain never to reach the posthuman stage or almost all posthuman civilizations lack individuals who run large numbers of ancestor-simulations, i.e. computer-emulations of the sort of human-like creatures from which they evolved [27]. This conclusion is a pessimistic one,

[14] For a brief summary of the Doomsday argument, see [74].

[15] This holds so long as the total number of Earth-like planets in the cosmos is sufficiently great to make it highly likely that at least some of them would develop intelligent observers [77].

for it narrows down quite substantially the range of positive future scenarios that are tenable in light of the empirical information we now have.

The Simulation argument does more than just sound a general alarm; it also redistributes probability among the hypotheses that remain believable. It increases the probability that we are living in a simulation (which may in many subtle ways affect our estimates of how likely various outcomes are) and it decreases the probability that the posthuman world would contain lots of free individuals who have large resources and human-like motives. This gives us some valuable hints as to what we may realistically hope for and consequently where we should direct our efforts.

8.5 Psychological biases?

The psychology of risk perception is an active but rather messy field [80] that could potentially contribute indirect grounds for reassessing our estimates of existential risks.

Suppose our intuitions about which future scenarios are "plausible and realistic" are shaped by what we see on TV and in movies and what we read in novels. (After all, a large part of the discourse about the future that people encounter is in the form of fiction and other recreational contexts.) We should then, when thinking critically, suspect our intuitions of being biased in the direction of overestimating the probability of those scenarios that make for a good story, since such scenarios will seem much more familiar and more "real". This *Good-story bias* could be quite powerful. When was the last time you saw a movie about humankind suddenly going extinct (without warning and without being replaced by some other civilization)? While this scenario may be much more probable than a scenario in which human heroes successfully repel an invasion of monsters or robot warriors, it wouldn't be much fun to watch. So we don't see many stories of that kind. If we are not careful, we can be mislead into believing that the boring scenario is too farfetched to be worth taking seriously. In general, if we think there is a Good-story bias, we may upon reflection want to increase our credence in boring hypotheses and decrease our credence in interesting, dramatic hypotheses. The net effect would be to redistribute probability among existential risks in favor of those that seem to harder to fit into a selling narrative, and possibly to increase the probability of the existential risks as a group.

The empirical data on risk-estimation biases is ambiguous. It has been argued that we suffer from various systematic biases when estimating our own prospects or risks in general. Some data suggest that humans tend to overestimate their own personal abilities and prospects.[16] About three quarters of all motorists think they are safer drivers than the

[16] Or at least that males do. One review [81] suggests that females underestimate their prospects although not by as much as males overestimate theirs. For more references, see [82], p. 489, [83,84].

typical driver.[17] Bias seems to be present even among highly educated people. According to one survey, almost half of all sociologists believed that they would become one of the top ten in their field [87], and 94% of sociologists thought they were better at their jobs than their average colleagues [88]. It has also been shown that depressives have a more accurate self-perception than normals except regarding the hopelessness of their situation [89-91]. Most people seem to think that they themselves are less likely to fall victims to common risks than other people [92]. It is widely believed [93] that the public tends to overestimate the probability of highly publicized risks (such as plane crashes, murders, food poisonings etc.), and a recent study [94] shows the public overestimating a large range of commonplace health risks to themselves. Another recent study [95], however, suggests that available data are consistent with the assumption that the public rationally estimates risk (although with a slight truncation bias due to cognitive costs of keeping in mind exact information).[18]

Even if we could get firm evidence for biases in estimating personal risks, we'd still have to be careful in making inferences to the case of existential risks.

8.6 Weighing up the evidence

In combination, these indirect arguments add important constraints to those we can glean from the direct consideration of various technological risks, although there is not room here to elaborate on the details. But the balance of evidence is such that it would appear unreasonable not to assign a substantial probability to the hypothesis that an existential disaster will do us in. My subjective opinion is that setting this probability lower than 25% would be misguided, and the best estimate may be considerably higher. But even if the probability were much smaller (say, ~1%) the subject matter would still merit very serious attention because of how much is at stake.

In general, the greatest existential risks on the time-scale of a couple of centuries or less appear to be those that derive from the activities of advanced technological civilizations. We see this by looking at the various existential risks we have listed. In each of the four categories, the top risks are engendered by our activities. The only significant existential risks for which this isn't true are "simulation gets shut down" (although on some versions of this hypothesis the shutdown would be prompted by our activities [27]); the catch-all hypotheses (which include both types of scenarios); asteroid

[17] For a review, see chapter 12 of [85]. Some of these studies neglect that it may well be *true* that 75% of drivers are better than the average driver; some studies, however, seem to avoid this problem, e.g. [86].

[18] Could the reason why recent studies speak more favorably about public rational risk assessment be that earlier results have resulted in public learning and recalibration? Researchers trying to establish systematic biases in risk perception could be shooting after a moving target much like those who attempt to find regularities in stock indexes. As soon as a consensus develops that there is such an effect, it disappears.

or comet impact (which is a very low probability risk); and getting killed by an extraterrestrial civilization (which would be highly unlikely in the near future).[19]

It may not be surprising that existential risks created by modern civilization get the lion's share of the probability. After all, we are now doing some things that have never been done on Earth before, and we are developing capacities to do many more such things. If non-anthropogenic factors have failed to annihilate the human species for hundreds of thousands of years, it could seem unlikely that such factors will strike us down in the next century or two. By contrast, we have no reason whatever not to think that the products of advanced civilization will be our bane.

We shouldn't be too quick to dismiss the existential risks that aren't human-generated as insignificant, however. It's true that our species has survived for a long time in spite of whatever such risks are present. But there may be an observation selection effect in play here. The question to ask is, on the theory that natural disasters sterilize Earth-like planets with a high frequency, what should we expect to observe? Clearly not that we are living on a sterilized planet. But maybe that we should be more primitive humans than we are? In order to answer this question, we need a solution to the problem of the reference class in observer selection theory [76]. Yet that is a part of the methodology that doesn't yet exist. So at the moment we can state that the most serious existential risks are generated by advanced human civilization, but we base this assertion on direct considerations. Whether there is additional support for it based on indirect considerations is an open question.

We should not *blame* civilization or technology for imposing big existential risks. Because of the way we have defined existential risks, a failure to develop technological civilization would imply that we had fallen victims of an existential disaster (namely a crunch, "technological arrest"). Without technology, our chances of avoiding existential risks would therefore be nil. With technology, we have some chance, although the greatest risks now turn out to be those generated by technology itself.

9 Implications for policy and ethics

Existential risks have a cluster of features that make it useful to identify them as a special category: the extreme magnitude of the harm that would come from an existential disaster; the futility of the trial-and-error approach; the lack of evolved biological and cultural coping methods; the fact that existential risk dilution is a global public good; the shared stakeholdership of all future generations; the international nature of many of the required countermeasures; the necessarily highly speculative and multidisciplinary nature of the topic; the subtle and diverse methodological problems involved in assessing the

[19] The crunch scenario "technological arrest" couldn't properly be said to be *caused* by our activities.

probability of existential risks; and the comparative neglect of the whole area. From our survey of the most important existential risks and their key attributes, we can extract tentative recommendations for ethics and policy:

9.1 Raise the profile of existential risks

We need more research into existential risks – detailed studies of particular aspects of specific risks as well as more general investigations of associated ethical, methodological, security and policy issues. Public awareness should also be built up so that constructive political debate about possible countermeasures becomes possible.

Now, it's a commonplace that researchers always conclude that more research needs to be done in their field. But in this instance it is *really* true. There is more scholarly work on the life-habits of the dung fly than on existential risks.

9.2 Create a framework for international action

Since existential risk reduction is a global public good, there should ideally be an institutional framework such that the cost and responsibility for providing such goods could be shared fairly by all people. Even if the costs can't be shared fairly, some system that leads to the provision of existential risk reduction in something approaching optimal amounts should be attempted.

The necessity for international action goes beyond the desirability of cost-sharing, however. Many existential risks simply cannot be substantially reduced by actions that are internal to one or even most countries. For example, even if a majority of countries pass and enforce national laws against the creation of some specific destructive version of nanotechnology, will we really have gained safety if some less scrupulous countries decide to forge ahead regardless? And strategic bargaining could make it infeasible to bribe all the irresponsible parties into subscribing to a treaty, even if everybody would be better off if everybody subscribed [14,42].

9.3 Retain a last-resort readiness for preemptive action

Creating a broad-based consensus among the world's nation states is time-consuming, difficult, and in many instances impossible. We must therefore recognize the possibility that cases may arise in which a powerful nation or a coalition of states needs to act unilaterally for its own and the common interest. Such unilateral action may infringe on the sovereignty of other nations and may need to be done preemptively.

Let us make this hypothetical more concrete. Suppose advanced nanotechnology has just been developed in some leading lab. (By advanced nanotechnology I mean a fairly general assembler, a device that can build a large range of three-dimensional structures – including rigid parts – to atomic precision given a detailed specification of the design and construction process, some feedstock chemicals, and a supply of energy.)

Suppose that at this stage it is possible to predict that building dangerous nanoreplicators will be much easier than building a reliable nanotechnological immune system that could protect against all simple dangerous replicators. Maybe design-plans for the dangerous replicators have already been produced by design-ahead efforts and are available on the Internet. Suppose furthermore that because most of the research leading up to the construction of the assembler, excluding only the last few stages, is available in the open literature; so that other laboratories in other parts of the world are soon likely to develop their own assemblers. What should be done?

With this setup, one can confidently predict that the dangerous technology will soon fall into the hands of "rogue nations", hate groups, and perhaps eventually lone psychopaths. Sooner or later somebody would then assemble and release a destructive nanobot and destroy the biosphere. The only option is to take action to prevent the proliferation of the assembler technology until such a time as reliable countermeasures to a nano-attack have been deployed.

Hopefully, most nations would be responsible enough to willingly subscribe to appropriate regulation of the assembler technology. The regulation would not need to be in the form of a ban on assemblers but it would have to limit temporarily but effectively the uses of assemblers, and it would have to be coupled to a thorough monitoring program. Some nations, however, may refuse to sign up. Such nations would first be pressured to join the coalition. If all efforts at persuasion fail, force or the threat of force would have to be used to get them to sign on.

A preemptive strike on a sovereign nation is not a move to be taken lightly, but in the extreme case we have outlined – where a failure to act would with high probability lead to existential catastrophe – it is a responsibility that must not be abrogated. Whatever moral prohibition there normally is against violating national sovereignty is overridden in this case by the necessity to prevent the destruction of humankind. Even if the nation in question has not yet initiated open violence, the mere decision to go forward with development of the hazardous technology in the absence of sufficient regulation must be interpreted as an act of aggression, for it puts the rest of the rest of the world at an even greater risk than would, say, firing off several nuclear missiles in random directions.

The intervention should be decisive enough to reduce the threat to an acceptable level but it should be no greater than is necessary to achieve this aim. It may even be appropriate to pay compensation to the people of the offending country, many of whom will bear little or no responsibility for the irresponsible actions of their leaders.

While we should hope that we are never placed in a situation where initiating force becomes necessary, it is crucial that we make room in our moral and strategic thinking for this contingency. Developing widespread recognition of the moral aspects of this scenario ahead of time is especially important, since without some degree of public support democracies will find it difficult to act decisively before there has been any

visible demonstration of what is at stake. Waiting for such a demonstration is decidedly not an option, because it might itself be the end.[20]

9.4 Differential technological development

If a feasible technology has large commercial potential, it is probably impossible to prevent it from being developed. At least in today's world, with lots of autonomous powers and relatively limited surveillance, and at least with technologies that do not rely on rare materials or large manufacturing plants, it would be exceedingly difficult to make a ban 100% watertight. For some technologies (say, ozone-destroying chemicals), imperfectly enforceable regulation may be all we need. But with other technologies, such as destructive nanobots that self-replicate in the natural environment, even a single breach could be terminal. The limited enforceability of technological bans restricts the set of feasible policies from which we can choose.

What we do have the power to affect (to what extent depends on how we define "we") is the *rate* of development of various technologies and potentially the *sequence* in which feasible technologies are developed and implemented. Our focus should be on what I want to call *differential technological development*: trying to retard the implementation of dangerous technologies and accelerate implementation of beneficial technologies, especially those that ameliorate the hazards posed by other technologies. In the case of nanotechnology, the desirable sequence would be that defense systems are deployed before offensive capabilities become available to many independent powers; for once a secret or a technology is shared by many, it becomes extremely hard to prevent further proliferation. In the case of biotechnology, we should seek to promote research into vaccines, anti-bacterial and anti-viral drugs, protective gear, sensors and diagnostics, and to delay as much as possible the development (and proliferation) of biological warfare agents and their vectors. Developments that advance offense and defense equally are neutral from a security perspective, unless done by countries we identify as responsible, in which case they are advantageous to the extent that they increase our technological superiority over our potential enemies. Such "neutral" developments can also be helpful in reducing the threat from natural hazards and they may of course also have benefits that are not directly related to global security.

Some technologies seem to be especially worth promoting because they can help in reducing a broad range of threats. Superintelligence is one of these. Although it has its own dangers (expounded in preceding sections), these are dangers that we will have to face at some point no matter what. But getting superintelligence early is desirable because it would help diminish other risks. A superintelligence could advise us on policy.

[20] The complexities of strategizing about the best way to prepare for nanotechnology become even greater when we take into account the possible memetic consequences of advocating various positions at various times. For some further reflections on managing the risks of nanotechnology, see [23,25,26,41,96-99].

Superintelligence would make the progress curve for nanotechnology much steeper, thus shortening the period of vulnerability between the development of dangerous nanoreplicators and the deployment of adequate defenses. By contrast, getting nanotechnology before superintelligence would do little to diminish the risks of superintelligence. The main possible exception to this is if we think that it is important that we get to superintelligence via uploading rather than through artificial intelligence. Nanotechnology would greatly facilitate uploading [39].

Other technologies that have a wide range of risk-reducing potential include intelligence augmentation, information technology, and surveillance. These can make us smarter individually and collectively, and can make it more feasible to enforce necessary regulation. A strong prima facie case therefore exists for pursuing these technologies as vigorously as possible.[21]

As mentioned, we can also identify developments outside technology that are beneficial in almost all scenarios. Peace and international cooperation are obviously worthy goals, as is cultivation of traditions that help democracies prosper.[22]

9.5 Support programs that directly reduce specific existential risks

Some of the lesser existential risks can be countered fairly cheaply. For example, there are organizations devoted to mapping potentially threatening near-Earth objects (e.g. NASA's Near Earth Asteroid Tracking Program, and the Space Guard Foundation). These could be given additional funding. To reduce the probability of a "physics disaster", a public watchdog could be appointed with authority to commission advance peer-review of potentially hazardous experiments. This is currently done on an ad hoc basis and often in a way that relies on the integrity of researchers who have a personal stake in the experiments going forth.

The existential risks of naturally occurring or genetically engineered pandemics would be reduced by the same measures that would help prevent and contain more limited epidemics. Thus, efforts in counter-terrorism, civil defense, epidemiological monitoring and reporting, developing and stockpiling antidotes, rehearsing emergency quarantine procedures, etc. could be intensified. Even abstracting from existential risks, it

[21] Of course, intelligence enhancements can make evil persons better at pursuing their wicked ambitions, and surveillance could be used by dictatorial regimes (and hammers can be used to crush skulls). Unmixed blessings are hard to come by. But on balance, these technologies still seem very worth promoting. In the case of surveillance, it seems important to aim for the two-way transparency advocated by David Brin [100], where we all can watch the agencies that watch us.

[22] With limited resources, however, it is crucial to prioritize wisely. A million dollars could currently make a vast difference to the amount of research done on existential risks; the same amount spent on furthering world peace would be like a drop in the ocean.

would probably be cost-effective to increase the fraction of defense budgets devoted to such programs.[23]

Reducing the risk of a nuclear Armageddon, whether accidental or intentional, is a well-recognized priority. There is a vast literature on the related strategic and political issues to which I have nothing to add here.

The longer-term dangers of nanotech proliferation or arms race between nanotechnic powers, as well as the whimper risk of "evolution into oblivion", may necessitate, even more than nuclear weapons, the creation and implementation of a coordinated global strategy. Recognizing these existential risks suggests that it is advisable to gradually shift the focus of security policy from seeking national security through unilateral strength to creating an integrated international security system that can prevent arms races and the proliferation of weapons of mass destruction. Which particular policies have the best chance of attaining this long-term goal is a question beyond the scope of this paper.

9.6 Maxipok: a rule of thumb for moral action

Previous sections have argued that the combined probability of the existential risks is very substantial. Although there is still a fairly broad range of differing estimates that responsible thinkers could make, it is nonetheless arguable that because the negative utility of an existential disaster is so enormous, the objective of reducing existential risks should be a dominant consideration when acting out of concern for humankind as a whole. It may be useful to adopt the following rule of thumb for moral action; we can call it *Maxipok*:

> *Maximize the probability of an okay outcome*, where an "okay outcome" is any outcome that avoids existential disaster.

At best, this is a rule of thumb, a prima facie suggestion, rather than a principle of absolute validity, since there clearly *are* other moral objectives than preventing terminal global disaster. Its usefulness consists in helping us to get our priorities straight. Moral action is always at risk to diffuse its efficacy on feel-good projects[24] rather on serious work that has the best chance of fixing the worst ills. The cleft between the feel-good projects and what really has the greatest potential for good is likely to be especially great in regard to existential risk. Since the goal is somewhat abstract and since existential risks

[23] This was written before the 9-11 tragedy. Since then, U.S. defense priories have shifted in the direction advocated here. I think still further shifts are advisable.
[24] See e.g. [101] and references therein.

don't currently cause suffering in any living creature[25], there is less of a feel-good dividend to be derived from efforts that seek to reduce them. This suggests an offshoot moral project, namely to reshape the popular moral perception so as to give more credit and social approbation to those who devote their time and resources to benefiting humankind via global safety compared to other philanthropies.

Maxipok, a kind of satisficing rule, is different from *Maximin* ("Choose the action that has the best worst-case outcome.")[26]. Since we cannot completely eliminate existential risks (at any moment we could be sent into the dustbin of cosmic history by the advancing front of a vacuum phase transition triggered in a remote galaxy a billion years ago) using maximin in the present context has the consequence that we should choose the act that has the greatest benefits under the assumption of impending extinction. In other words, maximin implies that we should all start partying as if there were no tomorrow.

While that option is indisputably attractive, it seems best to acknowledge that there just might be a tomorrow, especially if we play our cards right.

10 Acknowledgments

I'm grateful for comments to Curt Adams, Amara Angelica, Brian Atkins, Milan Cirkovic, Douglas Chamberlain, Robert A. Freitas Jr., Mark Gubrud, Robin Hanson, Barbara Lamar, John Leslie, Mike Treder, Ken Olum, Robert Pisani, several anonymous referees, and to the audience at a SIG meeting at the Foresight Institute's Senior Associates Gathering, April 2001, Palo Alto, where an earlier version of this paper was presented. The paper has also benefited from discussions with Michaela Fistioc, Bill Joy, John Oh, Pat Parker, Keith DeRose, and Peter Singer.

11 Appendix: The outline of an evolutionary whimper

This appendix outlines why there is a risk that we may end in an evolutionary whimper. The following eleven-links chain of reasoning is not intended to be a rigorous proof of any kind but rather something like a suggestive narrative minus literary embellishments. (For a fuller discussion of some of these ideas, see [61].)

---

[25] An exception to this is if we think that a large part of what's possible and desirable about a posthuman future is that it contains a large portion of the people who are currently alive. If take this view then the current global death rate of 150,000 persons/day is an aspect of an ongoing, potentially existential, disaster (a shriek) that is causing vast human suffering.

[26] Following John Rawls [102], the term "maximin" is also use in a different sense in welfare economics, to denote the principle that (given some important constraints) we should opt for the state that optimizes the expectation of the least well-off classes. This version of the principle is not necessarily affected by the remarks that follow.

1. Although it's easy to think of evolution as leading from simple to more complex life forms, we should not uncritically assume that this is always so. It is true that here on Earth, simple replicators have evolved to human beings (among other things), but because of an observation selection effect the evidential value of this single data point is very limited (more on this in the section on estimating the probability of existential risks).
2. We don't currently *see* much evolutionary development in the human species. This is because biological evolution operates on a time-scale of many generations, not because it doesn't occur any longer [103].
3. Biological human evolution is slow primarily because of the slowness of human reproduction (with a minimum generational lag of about one and a half decade).
4. Uploads and machine intelligences can reproduce virtually instantaneously, provided easy resources are available. Also, if they can predict some aspects of their evolution, they can modify themselves accordingly right away rather than waiting to be outcompeted. Both these factors can lead to a much more rapid evolutionary development in a posthuman world.
5. The activities and ways of being to which we attach value may not coincide with the activities that have the highest economic value in the posthuman world. Agents who choose to devote some fraction of their resources to (unproductive or less-than-optimally productive) "hobbies" would be at a competitive disadvantage, and would therefore risk being outcompeted. (So how could play evolve in humans and other primates? Presumably because it was adaptive and hence "productive" in the sense of the word used here. We place a value on play. But the danger consists in there being no guarantee that the activities that are adaptive in the future will be ones that we would currently regard as valuable – the adaptive activities of the future may not even be associated with any consciousness.)
6. We need to distinguish between two senses of "outcompeted". In the first sense, an outcompeted type is outcompeted only in a relative sense: the resources it possesses constitutes a smaller and smaller fraction of the total of colonized resources as time passes. In the second sense, an outcompeted type's possessions decrease in absolute terms so that the type eventually becomes extinct.
7. If property rights were nearly perfectly enforced (over cosmic distances, which seems hard to do) then the "hobbyists" (those types that devote some of their resources on activities that are unproductive) would be outcompeted only in the first sense. Depending on the details, this may or may not qualify as a whimper. If the lost potential (due to the increasing dominance of types that we don't regard as valuable) were great enough, it would be a whimper.

8. Without nearly perfect enforcement of property rights, we would have to fear that the hobbyists would become extinct because they are less efficient competitors for the same ecological niche than those types which don't expend any of their resources on hobbyist activities.
9. The only way of avoiding this outcome may be to replace natural evolution with *directed evolution*, i.e. by shaping the social selection pressures so that they favor the hobbyist type (by, for example, taxing the non-hobbyists) [19,104]. This could make the hobbyist type competitive.
10. Directed evolution, however, requires coordination. It is no good if some societies decide to favor their hobbyists if there are other societies that instead decide to maximize their productivity by not spending anything on subsidizing hobbyists. For the latter would then eventually outcompete the former. Therefore, the only way that directed evolution could avoid what would otherwise be a fated evolutionary whimper may be if there is on the highest level of organization only one independent agent. We can call such an organization a *singleton*.
11. A singleton does not need to be a monolith. It can contain within itself a highly diverse ecology of independent groups and individuals. A singleton could for example be a democratic world government or a friendly superintelligence [35]. Yet, whether a singleton will eventually form is an open question. If a singleton is not formed, and if the fitness landscape of future evolution doesn't favor dispositions to engage in activities we find valuable, then an evolutionary whimper may be the result.

12 References

1. Hanson, R. (1995). Could Gambling Save Science? Encouraging an Honest Consensus. *Social Epistemology, 9:1*, 3-33.

2. Tickner, J. et al. (2000). *The Precautionary Principle*. URL: http://www.biotech-info.net/handbook.pdf.

3. Foster, K.R. et al. (2000). Science and the Precautionary Principle. *Science, 288*, 979-981. URL: http://www.biotech-info.net/science_and_PP.html.

4. Lewis, D. (1986). *Philosophical Papers* (Vol. 2). New York: Oxford University Press.

5. Lewis, D. (1994). Humean Supervenience Debugged. *Mind, 103*(412), 473-490.

6. Bostrom, N. (1999). A Subjectivist Theory of Objective Chance, *British Society for the Philosophy of Science Conference, July 8-9*, Nottingham, U.K.

7. Jeffrey, R. (1965). *The logic of decision*: McGraw-Hill.

8. Kennedy, R. (1968). *13 Days*. London: Macmillan.

9. Leslie, J. (1996). *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

10. Putnam, H. (1979). The place of facts in a world of values. In D. Huff & O. Prewett (Eds.), *The Nature of the Physical Universe* (pp. 113-140). New York: John Wiley.

11. Kubrick, S. (1964). *Dr. Strangelove or How I Learned to Stop Worrying and Love the Bomb*: Columbia/Tristar Studios.

12. Shute, N. (1989). *On the Beach*: Ballentine Books.

13. Kaul, I. (1999). *Global Public Goods*: Oxford University Press.

14. Feldman, A. (1980). *Welfare Economics and Social Choice Theory*. Boston: Martinus Nijhoff Publishing.

15. Caplin, A., & Leahy, J. (2000). The Social Discount Rate. *National Bureau of Economic Research, Working paper 7983*.

16. Schelling, T.C. (2000). Intergenerational and International Discounting. *Risk Analysis, 20*(6), 833-837.

17. Earman, J. (1995). *Bangs, Crunches, Whimpers, and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*: Oxford University Press.

18. Morgan, M.G. (2000). Categorizing Risks for Risk Ranking. *Risk Analysis, 20*(1), 49-58.

19. Bostrom, N. et al. (1999). The Transhumanist FAQ. URL: http://www.transhumanist.org.

20. Powell, C. (2000). 20 Ways the World Could End. *Discover, 21*(10). URL: http://www.discover.com/oct_00/featworld.html.

21. Joy, B. (2000). Why the future doesn't need us. *Wired, 8.04*. URL: http://www.wired.com/wired/archive/8.04/joy_pr.html.

22. Drexler, K.E. (1992). *Nanosystems*. New York: John Wiley & Sons, Inc.

23. Drexler, K.E. (1985). *Engines of Creation: The Coming Era of Nanotechnology*. London: Forth Estate. URL: http://www.foresight.org/EOC/index.html.

24. Merkle, R. et al. (1991). Theoretical studies of a hydrogen abstraction tool for nanotechnology. *Nanotechnology, 2*, 187-195.

25. Freitas (Jr.), R.A. (2000). Some Limits to Global Ecophagy by Biovorous Nanoreplicators, with Public Policy Recommendations. *Zyvex preprint, April 2000*. URL: http://www.foresight.org/NanoRev/Ecophagy.html.

26. Gubrud, M. (2000). Nanotechnology and International Security, *Fifth Foresight Conference on Molecular Nanotechnology*. URL: http://www.foresight.org/Conferences/MNT05/Papers/Gubrud/index.html.

27. Bostrom, N. (2001). Are You Living in a Simulation? *Working-paper*. URL: http://www.simulation-argument.com.

28. Moravec, H. (1989). *Mind Children*. Harvard: Harvard University Press.

29. Moravec, H. (1999). *Robot: Mere Machine to Transcendent Mind*. New York: Oxford University Press.

30. Vinge, V. (1993). The Coming Technological Singularity. *Whole Earth Review, Winter issue*.

31. Bostrom, N. (1998). How Long Before Superintelligence? *International Journal of Futures Studies, 2*. URL: http://www.nickbostrom.com/superintelligence.html.

32. Moravec, H. (1998). When will computer hardware match the human brain? *Journal of Transhumanism, 1*. URL: http://www.transhumanist.com/volume1/moravec.htm.

33. Kurzweil, R. (1999). *The Age of Spiritual Machines: When computers exceed human intelligence*. New York: Viking.

34. Hanson, R. et al. (1998). A Critical Discussion of Vinge's Singularity Concept. *Extropy Online*. URL: http://www.extropy.org/eo/articles/vi.html.

35. Yudkowsky, E. (2001). Friendly AI 0.9. URL: http://singinst.org/CaTAI/friendly/contents.html.

36. National Intelligence Council (2000). Global Trends 2015: A Dialogue about the Future with Nongovernment Experts. URL: http://www.cia.gov/cia/publications/globaltrends2015/.

37. Nowak, R. (2001). Disaster in the making. *New Scientist, 13 January 2001*. URL: http://www.newscientist.com/nsplus/insight/bioterrorism/disasterin.html.

38. Jackson, R.J. et al. (2001). Expression of Mouse Interleukin-4 by a Recombinant Ectromelia Virus Suppresses Cytolytic Lymphocyte Responses and Overcomes Genetic Resistance to Mousepox. *Journal of Virology, 73*, 1479-1491.

39. Freitas (Jr.), R.A. (1999). *Nanomedicine, Volume 1: Basic Capabilities*. Georgetown, TX: Landes Bioscience. URL: http://www.nanomedicine.com.

40. Foresight Institute (2000). Foresight Guidelines on Molecular Nanotechnology. , *Version 3.7*. URL: http://www.foresight.org/guidelines/current.html.

41. Foresight Institute (1997-1991). Accidents, Malice, Progress, and Other Topics. *Background 2, Rev. 1*. URL: http://www.foresight.org/Updates/Background2.html.

42. Schelling, T.C. (1960). *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.

43. Knight, L.U. (2001). The Voluntary Human Extinction Movement. URL: http://www.vhemt.org/.

44. Schopenhauer, A. (1891). *Die Welt als Wille und Vorstellung*. Leipzig: F, A, Brockhaus.

45. Coleman, S., & Luccia, F. (1980). Gravitational effects on and of vacuum decay. *Physical Review D, 21*, 3305-3315.

46. Dar, A. et al. (1999). Will relativistic heavy-ion colliders destroy our planet? *Physics Letters, B 470*, 142-148.

47. Turner, M.S., & Wilczek, F. (1982). Is our vacuum metastable? *Nature, August 12*, 633-634.

48. Morrison, D. et al. (1994). The Impact Hazard. In T. Gehrels (Ed.), *Hazards Due to Comets and Asteroids*. Tucson: The University of Arizona Press.

49. Gold, R.E. (1999). SHIELD: A Comprehensive Earth Protection System. *A Phase I Report on the NASA Institute for Advanced Concepts, May 28, 1999*.

50. Huxley, A. (1932). *Brave New World*. London: Chatto & Windus.

51. Flynn, J.R. (1987). Massive IQ gains in many countries: What IQ tests really measure. *Psychological Bulletin, 101*, 171-191.

52. Storfer, M. (1999). Myopia, Intelligence, and the Expanding Human Neocortex. *International Journal of Neuroscience, 98*(3-4).

53. Merkle, R. (1994). The Molecular Repair of the Brain. *Cryonics, 15*(1 and 2).

54. Hanson, R. (1994). What If Uploads Come First: The crack of a future dawn. *Extropy, 6*(2). URL: http://hanson.gmu.edu/uploads.html.

55. Warwick, K. (1997). *March of the Machines*. London: Century.

56. Whitby, B. et al. (2000). How to Avoid a Robot Takeover: Political and Ethical Choices in the Design and Introduction of Intelligent Artifacts. *Presented at AISB-00 Symposium on Artificial Intelligence, Ethics an (Quasi-) Human Rights*. URL: http://www.cogs.susx.ac.uk/users/blayw/BlayAISB00.html.

57. Ettinger, R. (1964). *The prospect of immortality*. New York: Doubleday.

58. Zehavi, I., & Dekel, A. (1999). Evidence for a positive cosmological constant from flows of galaxies and distant supernovae. *Nature, 401*(6750), 252-254.

59. Bostrom, N. (2001). Are Cosmological Theories Compatible With All Possible Evidence? A Missing Methodological Link. *In preparation*.

60. Cirkovic, M., & Bostrom, N. (2000). Cosmological Constant and the Final Anthropic Hypothesis. *Astrophysics and Space Science, 274*(4), 675-687. URL: http://xxx.lanl.gov.

61. Bostrom, N. (2001). The Future of Human Evolution. *Working paper*. URL: http://www.nickbostrom.com.

62. Hanson, R. (1998). Burning the Cosmic Commons: Evolutionary Strategies for Interstellar Colonization. *Working paper*. URL: http://hanson.gmu.edu/workingpapers.html.

63. Freitas(Jr.), R.A. (1980). A Self-Reproducing Interstellar Probe. *J. Brit. Interplanet. Soc., 33*, 251-264.

64. Bostrom, N. (2000). Predictions from Philosophy? *Coloquia Manilana (PDCIS), 7*. URL: http://www.nickbostrom.com/old/predict.html.

65. Chislenko, A. (1996). Networking in the Mind Age. URL: http://www.lucifer.com/~sasha/mindage.html.

66. Barrow, J.D., & Tipler, F.J. (1986). *The Anthropic Cosmological Principle*. Oxford: Oxford University Press.

67. Tipler, F.J. (1982). Anthropic-principle arguments against steady-state cosmological theories. *Observatory, 102*, 36-39.

68. Brin, G.D. (1983). The 'Great Silence': The Controversy Concerning Extraterrestrial Intelligent Life. *Quarterly Journal of the Royal Astronomical Society, 24*, 283-309.

69. Hanson, R. (1998). The Great Filter - Are We Almost Past It? *Working paper.*

70. Carter, B. (1983). The anthropic principle and its implications for biological evolution. *Phil. Trans. R. Soc., A 310*, 347-363.

71. Carter, B. (1989). The anthropic selection principle and the ultra-Darwinian synthesis. In F. Bertola & U. Curi (Eds.), *The anthropic principle* (pp. 33-63). Cambridge: Cambridge University Press.

72. Hanson, R. (1998). Must Early Life be Easy? The rhythm of major evolutionary transitions. URL: http://hanson.berkeley.edu/.

73. Leslie, J. (1989). Risking the World's End. *Bulletin of the Canadian Nuclear Society, May, 10-15*.

74. Bostrom, N. (2000). Is the end nigh?, *The philosopher's magazine, Vol. 9* (pp. 19-20). URL: http://www.anthropic-principle.com/primer.html.

75. Bostrom, N. (1999). The Doomsday Argument is Alive and Kicking. *Mind, 108*(431), 539-550. URL: http://www.anthropic-principle.com/preprints/ali/alive.html.

76. Bostrom, N. (2002). *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. Routledge, New York. URL: http://www.anthropic-principle.com/book/.

77. Bostrom, N. (2001). Fine-Tuning Arguments in Cosmology. *In preparation*. URL: http://www.anthropic-principle.com.

78. Bostrom, N. (2000). Observer-relative chances in anthropic reasoning? *Erkenntnis, 52*, 93-108. URL: http://www.anthropic-principle.com/preprints.html.

79. Bostrom, N. (2001). The Doomsday argument, Adam & Eve, UN++, and Quantum Joe. *Synthese, 127*(3), 359-387. URL: http://www.anthropic-principle.com.

80. Sjöberg, L. (2000). Factors in Risk Perception. *Risk Analysis, 20*(1), 1-11.

81. Frieze, I. et al. (1978). *Women and sex roles*. New York: Norton.

82. Waldeman, M. (1994). Systematic Errors and the Theory of Natural Selection. *The American Economics Review, 84*(3), 482-497.

83. Cowen, T., & Hanson, R. (2001). How YOU Do Not Tell the Truth: Academic Disagreement as Self-Deception. *Working paper*.

84. Kruger, J., & Dunning, D. (1999). Unskilled and Unaware if It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments. *Journal of Personality and Social Psychology, 77*(6), 1121-1134.

85. Evans, L. (1991). *Traffic Safety and the Driver*: Leonard Evans. URL: http://www.scienceservingsociety.com/book/.

86. Svenson, O. (1981). Are we less risky and more skillful that our fellow drivers? *Acta Psychologica, 47*, 143-148.

87. Westie, F.R. (1973). Academic Expectations of Professional Immortality: A Study of Legitimation. *The American Sociologists, 8*, 19-32.

88. Gilovich, T. (1991). *How We Know What Isn't So*. New York: Macmillan.

89. Paulhaus, D.L. (1986). Self-Deception and Impression Management in Test Responses. In A. Angeitner & J.S. Wiggins (Eds.), *Personality Assessment via Questionnaires: Current Issues in Theory and Measurement*. New York: Springer.

90. Roth, D.L., & Ingram, R.E. (1985). Factors in the Self-Deception Questionnaire: Associations with depression. *Journal of Personality and Social Psychology, 48*, 243-251.

91. Sackheim, H.A., & Gur, R.C. (1979). Self-deception, other-deception, and self-reported psychopathology. *Journal of Consulting and Clinical Psychology, 47*, 213-215.

92. Sjöberg, L. (1994). Stralforskningens Risker: Attityder, Kunskaper och Riskuppfattning. *RHIZIKON: Rapport fran Centrum för Riskforskning, Handelshögskolan i Stockholm, 1*.

93. Urguhart, J., & Heilmann, K. (1984). *Risk Watch: The Odds of Life*. New York: Facts on File Publications.

94. Taylor, H. (1999). Perceptions of Risks. *The Harris Poll #7, January 27*. URL: http://www.harrisinteractive.com/harris_poll/index.asp?PID=44.

95. Benjamin, D.K. et al. (2001). Individuals' estimates of risks of death: Part II - New evidence. *Journal of Risk and Uncertainty, 22*(1), 35-57.

96. Drexler, K.E. (1988). A Dialog on Dangers. *Foresight Background 2, Rev. 1*. URL: http://www.foresight.org/Updates/Background3.html.

97. McCarthy, T. (2000). Molecular Nanotechnology and the World System. . URL: http://www.mccarthy.cx/WorldSystem/intro.htm.

98. Forrest, D. (1989). Regulating Nanotechnology Development. . URL: http://www.foresight.org/NanoRev/Forrest1989.html.

99. Jeremiah, D.E. (1995). Nanotechnology and Global Security. *Presented at the Fourth Foresight Conference on Molecular Nanotechnology*. URL: http://www.zyvex.com/nanotech/nano4/jeremiahPaper.html.

100. Brin, D. (1998). *The Transparent Society*. Reading, MA.: Addison-Wesley.

101. Hanson, R. (2000). Showing That You Care: The Evolution of Health Altruism. . URL: http://hanson.gmu.edu/bioerr.pdf.

102. Rawls, J. (1999). *A Theory of Justice* (Revised Edition ed.). Cambridge, Mass.: Harvard University Press.

103. Kirk, K.M. (2001). Natural Selection and Quantitative Genetics of Life-History Traits in Western Women: A Twin Study. *Evolution, 55*(2), 432-435. URL: http://evol.allenpress.com/evolonline/?request=get-document&issn=0014-3820&volume=055&issue=02&page=0423.

104. Bostrom, N. (2001). Transhumanist Values. *Manuscript*. URL: http://www.nickbostrom.com.