# The Future of Humanity

Nick Bostrom
Oxford University
www.nickbostrom.com

Complete draft circulated (2007). Published in *New Waves in Philosophy of Technology*, eds. Jan-Kyrre Berg Olsen, Evan Selinger, & Soren Riis (New York: Palgrave Macmillan, 2009): 186-215 Reprinted in the journal *Geopolitics, History, and International Relations*, Vol. 1, No. 2 (2009): 41-78.

Formats: HTML, PDF

**Abstract**

The future of humanity is often viewed as a topic for idle speculation. Yet our beliefs and assumptions on this subject matter shape decisions in both our personal lives and public policy – decisions that have very real and sometimes unfortunate consequences. It is therefore practically important to try to develop a realistic mode of futuristic thought about big picture questions for humanity. This paper sketches an overview of some recent attempts in this direction, and it offers a brief discussion of four families of scenarios for humanity's future: extinction, recurrent collapse, plateau, and posthumanity.

---

## The future of humanity as an inescapable topic

In one sense, the future of humanity comprises everything that will ever happen to any human being, including what you will have for breakfast next Thursday and all the scientific discoveries that will be made next year. In that sense, it is hardly reasonable to think of the future of humanity as a *topic*: it is too big and too diverse to be addressed as a whole in a single essay, monograph, or even 100-volume book series. It is made into a topic by way of abstraction. We abstract from details and short-term fluctuations and developments that affect only some limited aspect of our lives. A discussion about the future of humanity is about how the important fundamental features of the human condition may change or remain constant in the long run.

What features of the human condition are fundamental and important? On this there can be reasonable disagreement. Nonetheless, some features qualify by almost any standard. For example, whether and when Earth-originating life will go extinct, whether it will colonize the galaxy, whether human biology will be fundamentally transformed to make us posthuman, whether machine intelligence will surpass biological intelligence, whether population size will explode, and whether quality of life will radically improve or deteriorate: these are all important fundamental questions about the future of humanity. Less fundamental questions – for instance, about methodologies or specific technology projections – are also relevant insofar as they inform our views about more fundamental parameters.

Traditionally, the future of humanity has been a topic for theology. All the major religions have teachings about the ultimate destiny of humanity or the end of the world.[1] Eschatological themes have also been explored by big-name philosophers such as Hegel, Kant, and Marx. In more recent times the literary genre of science fiction has continued the tradition. Very often, the future has served as a projection screen for our hopes and fears; or as a stage setting for dramatic entertainment, morality tales, or satire of tendencies in contemporary society; or as a banner for ideological mobilization. It is relatively rare for humanity's future to be taken seriously as a subject matter on which it is important to try to have factually correct beliefs. There is nothing wrong with exploiting the symbolic and literary affordances of an unknown future, just as there is nothing wrong with fantasizing about imaginary countries populated by dragons and wizards. Yet it is important to attempt (as best we can) to distinguish futuristic scenarios put forward for their symbolic significance or entertainment value from speculations that are meant to be evaluated on the basis of literal plausibility. Only the latter form of "realistic" futuristic thought will be considered in this paper.

We need realistic pictures of what the future might bring in order to make sound decisions. Increasingly, we need realistic pictures not only of our personal or local near-term futures, but also of remoter global futures. Because of our expanded technological powers, some human activities now have significant global impacts. The scale of human social organization has also grown, creating new opportunities for coordination and action, and there are many institutions and individuals who either *do* consider, or *claim* to consider, or *ought* to consider, possible long-term global impacts of their actions. Climate change, national and international security, economic development, nuclear waste disposal, biodiversity, natural resource conservation, population policy, and scientific and technological research funding are examples of policy areas that involve long time-horizons. Arguments in these areas often rely on implicit assumptions about the future of humanity. By making these assumptions explicit, and subjecting them to critical analysis, it might be possible to address some of the big challenges for humanity in a more well-considered and

1 (Hughes 2007)

thoughtful manner.

The fact that we "need" realistic pictures of the future does not entail that we can have them. Predictions about future technical and social developments are notoriously unreliable – to an extent that has led some to propose that we do away with prediction altogether in our planning and preparation for the future. Yet while the methodological problems of such forecasting are certainly very significant, the extreme view that we can or should do away with prediction altogether is misguided. That view is expressed, to take one example, in a recent paper on the societal implications of nanotechnology by Michael Crow and Daniel Sarewitz, in which they argue that the issue of predictability is "irrelevant":

> preparation for the future obviously does not require accurate prediction; rather, it requires a foundation of knowledge upon which to base action, a capacity to learn from experience, close attention to what is going on in the present, and healthy and resilient institutions that can effectively respond or adapt to change in a timely manner.[2]

Note that each of the elements Crow and Sarewitz mention as required for the preparation for the future relies in some way on accurate prediction. A capacity to learn from experience is not useful for preparing for the future unless we can correctly assume (predict) that the lessons we derive from the past will be applicable to future situations. Close attention to what is going on in the present is likewise futile unless we can assume that what is going on in the present will reveal stable trends or otherwise shed light on what is likely to happen next. It also requires non-trivial prediction to figure out what kind of institution will prove healthy, resilient, and effective in responding or adapting to future changes.

The reality is that predictability is a matter of degree, and different aspects of the future are predictable with varying degrees of reliability and precision.[3] It may often be a good idea to develop plans that are flexible and to pursue policies that are robust under a wide range of contingencies. In some cases, it also makes sense to adopt a reactive approach that relies on adapting quickly to changing circumstances rather than pursuing any detailed long-term plan or explicit agenda. Yet these coping strategies are only one part of the solution. Another part is to work to improve the accuracy of our beliefs about the future (including the accuracy of conditional predictions of the form "if x is done, y will result"). There might be traps that we are walking towards that we could only avoid falling into by means of foresight. There are also opportunities that we could reach much sooner if we could see them farther in advance. And in a

2 (Crow and Sarewitz 2001)

3 For example, it is likely that computers will become faster, materials will become stronger, and medicine will cure more diseases; cf. (Drexler 2003).

strict sense, prediction is *always* necessary for meaningful decision-making.[4]

Predictability does not necessarily fall off with temporal distance. It may be highly unpredictable where a traveler will be one hour after the start of her journey, yet predictable that after five hours she will be at her destination. The *very* long-term future of humanity may be relatively easy to predict, being a matter amenable to study by the natural sciences, particularly cosmology (physical eschatology). And for there to be a degree of predictability, it is not necessary that it be possible to identify one specific scenario as what will definitely happen. If there is at least some scenario that can be *ruled out*, that is also a degree of predictability. Even short of this, if there is some basis for assigning different probabilities (in the sense of credences, degrees of belief) to different propositions about logically possible future events, or some basis for criticizing some such probability distributions as less rationally defensible or reasonable than others, then again there is a degree of predictability. And this is surely the case with regard to many aspects of the future of humanity. While our knowledge is insufficient to narrow down the space of possibilities to one broadly outlined future for humanity, we do know of many relevant arguments and considerations which in combination impose significant constraints on what a plausible view of the future could look like. The future of humanity need not be a topic on which all assumptions are entirely arbitrary and anything goes. There is a vast gulf between knowing exactly what will happen and having absolutely no clue about what will happen. Our actual epistemic location is some offshore place in that gulf.[5]

## Technology, growth, and directionality

Most differences between our lives and the lives of our hunter-gatherer forebears are ultimately tied to technology, especially if we understand "technology" in its broadest sense, to include not only gadgets and machines but also techniques, processes, and institutions. In this wide sense we could say that technology is the sum total of instrumentally useful culturally-transmissible information. Language is a technology in this sense, along with tractors, machine guns, sorting algorithms, double-entry bookkeeping, and Robert's Rules of Order.[6]

Technological innovation is the main driver of long-term economic growth. Over long time scales, the compound effects of even modest average annual growth are profound. Technological change is in large part responsible for many of

4 You lift the glass to your mouth because you predict that drinking will quench your thirst; you avoid stepping in front of a speeding car because you predict that a collision will hurt you.

5 For more on technology and uncertainty, see (Bostrom 2007b).

6 I'm cutting myself some verbal slack. On the proposed terminology, a particular physical object such as farmer Bob's tractor is not, strictly speaking, technology but rather a *technological artifact*, which depends on and embodies technology-as-information. The individual tractor is physical capital. The transmissible information needed to produce tractors is technology.

the secular trends in such basic parameters of the human condition as the size of the world population, life expectancy, education levels, material standards of living, and the nature of work, communication, health care, war, and the effects of human activities on the natural environment. Other aspects of society and our individual lives are also influenced by technology in many direct and indirect ways, including governance, entertainment, human relationships, and our views on morality, mind, matter, and our own human nature. One does not have to embrace any strong form of technological determinism to recognize that technological capability – through its complex interactions with individuals, institutions, cultures, and environment – is a key determinant of the ground rules within which the games of human civilization get played out.[7]

This view of the important role of technology is consistent with large variations and fluctuations in deployment of technology in different times and parts of the world. The view is also consistent with technological development itself being dependent on socio-cultural, economic, or personalistic enabling factors. The view is also consistent with denying any strong version of inevitability of the particular growth pattern observed in human history. One might hold, for example, that in a "re-run" of human history, the timing and location of the Industrial Revolution might have been very different, or that there might not have been any such revolution at all but rather, say, a slow and steady trickle of invention. One might even hold that there are important bifurcation points in technological development at which history could take either path with quite different results in what kinds of technological systems developed. Nevertheless, *under the assumption that technological development continues on a broad front*, one might expect that *in the long run*, most of the important basic capabilities that could be obtained through some possible technology, will in fact be obtained through technology. A bolder version of this idea could be formulated as follows:

> *Technological Completion Conjecture.* If scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained.

The conjecture is not tautological. It would be false if there is some possible basic capability that could be obtained through some technology which, while possible in the sense of being consistent with physical laws and material constraints, is so difficult to develop that it would remain beyond reach even after an indefinitely prolonged development effort. Another way in which the conjecture could be false is if some important capability can only be achieved through some possible technology which, while it could have been developed, will not in fact ever be developed even though scientific and technological development efforts continue.

The conjecture expresses the idea that which important basic capabilities are eventually attained does not depend on the paths taken by scientific and

7 See e.g. (Wright 1999).

technological research in the short term. The principle allows that we might attain some capabilities sooner if, for example, we direct research funding one way rather than another; but it maintains that provided our general techno-scientific enterprise continues, even the non-prioritized capabilities will eventually be obtained, either through some indirect technological route, or when general advancements in instrumentation and understanding have made the originally neglected direct technological route so easy that even a tiny effort will succeed in developing the technology in question.[8]

One might find the thrust of this underlying idea plausible without being persuaded that the Technological Completion Conjecture is strictly true, and in that case, one may explore what exceptions there might be. Alternatively, one might accept the conjecture but believe that its antecedent is false, i.e. that scientific and technological development efforts will at some point effectively cease (before the enterprise is complete). But if one accepts both the conjecture and its antecedent, what are the implications? What will be the results if, in the long run, all of the important basic capabilities that could be obtained through some possible technology are in fact obtained? The answer may depend on the order in which technologies are developed, the social, legal, and cultural frameworks within which they are deployed, the choices of individuals and institutions, and other factors, including chance events. The obtainment of a basic capability does not imply that the capability will be used in a particular way or even that it will be used at all.

These factors determining the uses and impacts of potential basic capabilities are often hard to predict. What might be somewhat more foreseeable is which important basic capabilities will eventually be attained. For under the assumption that the Technological Completion Conjecture and its antecedent are true, the capabilities that will eventually be attained include all the ones that could be obtained through some possible technology. While we may not be able to foresee all possible technologies, we can foresee many possible technologies, including some that are currently infeasible; and we can show that these anticipated possible technologies would provide a large range of new important basic capabilities.

One way to foresee possible future technologies is through what Eric Drexler has termed "theoretical applied science".[9] Theoretical applied science studies the properties of possible physical systems, including ones that cannot yet be built, using methods such as computer simulation and derivation from

8 For a visual analogy, picture a box with large but finite volume, representing the space of basic capabilities that could be obtained through some possible technology. Imagine sand being poured into this box, representing research effort. The way in which you pour the sand will determine the places and speed at which piles build up in the box. Yet if you keep pouring, eventually the whole space gets filled.

9 (Drexler 1992)

established physical laws.[10] Theoretical applied science will not in every instance deliver a definitive and uncontroversial yes-or-no answer to questions about the feasibility of some imaginable technology, but it is arguably the best method we have for answering such questions. Theoretical applied science – both in its more rigorous and its more speculative applications – is therefore an important methodological tool for thinking about the future of technology and, a fortiori, one key determinant of the future of humanity.

It may be tempting to refer to the expansion of technological capacities as "progress". But this term has evaluative connotations – of things getting better – and it is far from a *conceptual* truth that expansion of technological capabilities makes things go better. Even if empirically we find that such an association has held in the past (no doubt with many big exceptions), we should not uncritically assume that the association will always continue to hold. It is preferable, therefore, to use a more neutral term, such as "technological development", to denote the historical trend of accumulating technological capability.

Technological development has provided human history with a kind of directionality. Instrumentally useful information has tended to accumulate from generation to generation, so that each new generation has begun from a different and technologically more advanced starting point than its predecessor. One can point to exceptions to this trend, regions that have stagnated or even regressed for extended periods of time. Yet looking at human history from our contemporary vantage point, the macro-pattern is unmistakable.

It was not always so. Technological development for most of human history was so slow as to be indiscernible. When technological development was that slow, it could only have been detected by comparing how levels of technological capability differed over large spans of time. Yet the data needed for such comparisons – detailed historical accounts, archeological excavations with carbon dating, and so forth – were unavailable until fairly recently, as Robert Heilbroner explains:

> At the very apex of the first stratified societies, dynastic dreams were dreamt and visions of triumph or ruin entertained; but there is no mention in the papyri and cuneiform tablets on which these hopes and fears were recorded that they envisaged, in the slightest degree, changes in the material conditions of the great masses, or for that matter, of the ruling class itself.[11]

Heilbroner argued in *Visions of the Future* for the bold thesis that humanity's perceptions of the shape of things to come have gone through exactly three phases since the first appearance of Homo sapiens. In the first phase, which comprises all of human prehistory and most of history, the worldly future

10 Theoretical applied science might also study potential pathways to the technology that would enable the construction of the systems in question, that is, how in principle one could solve the bootstrap problem of how to get from here to there.

11 (Heilbroner 1995), p. 8

was envisaged – with very few exceptions – as changeless in its material, technological, and economic conditions. In the second phase, lasting roughly from the beginning of the eighteenth century until the second half of the twentieth, worldly expectations in the industrialized world changed to incorporate the belief that the hitherto untamable forces of nature could be controlled through the appliance of science and rationality, and the future became a great beckoning prospect. The third phase – mostly post-war but overlapping with the second phase – sees the future in a more ambivalent light: as dominated by impersonal forces, as disruptive, hazardous, and foreboding as well as promising.

Supposing that some perceptive observer in the past had noticed some instance of directionality – be it a technological, cultural, or social trend – the question would have remained whether the detected directionality was a global feature or a mere local pattern. In a cyclical view of history, for example, there can be long stretches of steady cumulative development of technology or other factors. Within a period, there is clear directionality; yet each flood of growth is followed by an ebb of decay, returning things to where they stood at the beginning of the cycle. Strong local directionality is thus compatible with the view that, globally, history moves in circles and never really gets anywhere. If the periodicity is assumed to go on forever, a form of eternal recurrence would follow.

Modern Westerners who are accustomed to viewing history as a directional pattern of development may not appreciate how natural the cyclical view of history once seemed.[12] Any closed system with only a finite number of possible states must either settle down into one state and remain in that one state forever, or else cycle back through states in which it has already been. In other words, a closed finite state system must either become static or else start repeating itself. If we assume that the system has already been around for an eternity, then this eventual outcome must already have come about; i.e., the system is already either stuck or is cycling through states in which it has been before. The proviso that the system has only a finite number of states may not be as significant as it seems, for even a system that has an infinite number of possible states may only have finitely many *perceptibly different* possible states.[13] For many practical purposes, it may not matter much whether the current state of the world has already occurred an infinite number of times, or whether an infinite number of states have previously occurred each of which is merely imperceptibly different

12 The cyclical pattern is prominent in dharmic religions. The ancient Mayans held a cyclical view, as did many in ancient Greece. In the more recent Western tradition, the thought of eternal recurrence is most strongly associated with Nietzsche's philosophy, but the idea has been explored by numerous thinkers and is a common trope in popular culture.

13 The proviso of a *closed* system may also not have seemed significant. The universe is a closed system. The universe may not be a finite state system, but any finite part of the universe may permit of only finitely many different configurations, or finitely many perceptibly different configurations, allowing a kind of recurrence argument. In the actual case, an analogous result may hold with regard to spatial rather than temporal repetition. If we are living in a "Big World" then all possible human observations are in fact made by some observer (in fact, by infinitely many observers); see (Bostrom 2002c).

from the present state.[14] Either way, we could characterize the situation as one of eternal recurrence – the extreme case of a cyclical history.

In the actual world, the cyclical view is false because the world had a beginning a finite time ago. The human species has existed for a mere two hundred thousand years or so, and this is far from enough time for it to have experienced all possible conditions and permutations of which the system of humans and their environment is capable.

More fundamentally, the reason why the cyclical view is false is that the universe itself has existed for only a finite amount of time.[15] The universe started with the Big Bang an estimated 13.7 billion years ago, in a low-entropy state. The history of the universe has its own directionality: an ineluctable increase in entropy. During its process of entropy increase, the universe has progressed through a sequence of distinct stages. In the eventful first three seconds, a number of transitions occurred, including probably a period of inflation, reheating, and symmetry breaking. These were followed, later, by nucleosynthesis, expansion, cooling, and formation of galaxies, stars, and planets, including Earth (circa 4.5 billion years ago). The oldest undisputed fossils are about 3.5 billion years old, but there is some evidence that life already existed 3.7 billion years ago and possibly earlier. Evolution of more complex organisms was a slow process. It took some 1.8 billion years for eukaryotic life to evolve from prokaryotes, and another 1.4 billion years before the first multicellular organisms arose. From the beginning of the Cambrian period (some 542 million years ago), "important developments" began happening at a faster pace, but still enormously slowly by human standards. Homo habilis – our first "human-like ancestors" – evolved some 2 million years ago; Homo sapiens 100,000 years ago. The agricultural revolution began in the Fertile Crescent of the Middle East 10,000 years ago, and the rest is history. The size of the human population, which was about 5 million when we were living as hunter-gatherers 10,000 years ago, had grown to about 200 million by the year 1; it reached one billion in 1835 AD; and today over 6.6 billion human beings are breathing on this planet.[16] From the time of the industrial revolution, perceptive individuals living in developed countries have noticed significant technological change within their lifetimes.

All techno-hype aside, it is striking how recent many of the events are that define what we take to be the modern human condition. If we compress the time scale such that the Earth formed one year ago, then Homo sapiens evolved less than 12 minutes ago, agriculture began a little over one minute ago, the Industrial Revolution took place less than 2 seconds ago, the electronic computer was

14 It could matter if one accepted the "Unification" thesis. For a definition of this thesis, and an argument against it, see (Bostrom 2006).

15 According to the consensus model; but for a dissenting view, see e.g. (Steinhardt and Turok 2002).

16 (Bureau 2007). There is considerable uncertainty about the numbers especially for the earlier dates.

invented 0.4 seconds ago, and the Internet less than 0.1 seconds ago – in the blink of an eye.

Almost all the volume of the universe is ultra-high vacuum, and almost all of the tiny material specks in this vacuum are so hot or so cold, so dense or so dilute, as to be utterly inhospitable to organic life. Spatially as well as temporally, our situation is an anomaly.[17]

Given the technocentric perspective adopted here, and in light of our incomplete but substantial knowledge of human history and its place in the universe, how might we structure our expectations of things to come? The remainder of this paper will outline four families of scenarios for humanity's future:

- Extinction
- Recurrent collapse
- Plateau
- Posthumanity

## Extinction

Unless the human species lasts literally forever, it will some time cease to exist. In that case, the long-term future of humanity is easy to describe: extinction. An estimated 99.9% of all species that ever existed on Earth are already extinct.[18]

There are two different ways in which the human species could become extinct: one, by evolving or developing or transforming into one or more new species or life forms, sufficiently different from what came before so as no longer to count as Homo sapiens; the other, by simply dying out, without any meaningful replacement or continuation. Of course, a transformed continuant of the human species might itself eventually terminate, and perhaps there will be a point where all life comes to an end; so scenarios involving the first type of extinction may eventually converge into the second kind of scenario of complete annihilation. We postpone discussion of transformation scenarios to a later section, and we shall not here discuss the possible existence of fundamental physical limitations to the survival of intelligent life in the universe. This section focuses on the direct form of extinction (annihilation) occurring within any very long, but not astronomically long, time horizon – we could say one hundred thousand years for specificity.

Human extinction risks have received less scholarly attention than they deserve. In recent years, there have been approximately three serious books and one major paper on this topic. John Leslie, a Canadian philosopher, puts the probability of humanity failing to survive the next five centuries to 30% in his

17 Does anything interesting follow from this observation? Well, it is connected to a number of issues that do matter a great deal to work on the future of humanity – issues like observation selection theory and the Fermi paradox; cmp. (Bostrom 2002a).

18 (Raup 1991), p. 3f.

book *End of the World*.[19] His estimate is partly based on the controversial "Doomsday argument" and on his own views about the limitations of this argument.[20] Sir Martin Rees, Britain's Astronomer Royal, is even more pessimistic, putting the odds that humanity will survive the 21st century to no better than 50% in *Our Final Hour*.[21] Richard Posner, an eminent American legal scholar, offers no numerical estimate but rates the risk of extinction "significant" in *Catastrophe*.[22] And I published a paper in 2002 in which I suggested that assigning a probability of less than 25% to existential disaster (no time limit) would be misguided.[23] The concept of *existential risk* is distinct from that of extinction risk. As I introduced the term, an existential disaster is one that causes either the annihilation of Earth-originating intelligent life or the permanent and drastic curtailment of its potential for future desirable development.[24]

It is possible that a publication bias is responsible for the alarming picture presented by these opinions. Scholars who believe that the threats to human survival are severe might be more likely to write books on the topic, making the threat of extinction seem greater than it really is. Nevertheless, it is noteworthy that there seems to be a consensus among those researchers who have seriously looked into the matter that there is a serious risk that humanity's journey will come to a premature end.[25]

The greatest extinction risks (and existential risks more generally) arise from human activity. Our species has survived volcanic eruptions, meteoric impacts, and other natural hazards for tens of thousands of years. It seems unlikely that any of these old risks should exterminate us in the near future. By contrast, human civilization is introducing many novel phenomena into the world, ranging from nuclear weapons to designer pathogens to high-energy particle colliders. The most severe existential risks of this century derive from expected

19 (Leslie 1996)

20 Leslie defends the Carter-Leslie Doomsday argument, which leads to a strong probability shift in favor of "doom" (i.e. human extinction) occurring sooner rather than later. Yet Leslie also believes that the force of the Doomsday argument is weakened by quantum indeterminacy. Both of these beliefs – that the Doomsday argument is sound, and that if it is sound its conclusion would be weakened by quantum indeterminacy – are highly controversial. For a critical assessment, see (Bostrom 2002a).

21 (Rees 2003)

22 (Posner 2004)

23 (Bostrom 2002b)

24 Some scenarios in which the human species goes extinct may not be existential disasters – for example, if by the time of the disappearance of Homo sapiens we have developed new forms of intelligent life that continue and expand on what we valued in old biological humanity. Conversely, not all existential disasters involve extinction. For example, a global tyranny, if it could never be overthrown and if it were sufficiently horrible, would constitute an existential disaster even if the human species continued to exist.

25 A recent popular article by Bill Joy has also done much to disseminate concern about extinction risks. Joy's article focuses on the risks from genetics, nanotechnology, and robotics (artificial intelligence); (Joy 2000).

technological developments. Advances in biotechnology might make it possible to design new viruses that combine the easy contagion and mutability of the influenza virus with the lethality of HIV. Molecular nanotechnology might make it possible to create weapons systems with a destructive power dwarfing that of both thermonuclear bombs and biowarfare agents.[26] Superintelligent machines might be built and their actions could determine the future of humanity – and whether there will be one.[27] Considering that many of the existential risks that now seem to be among the most significant were conceptualized only in recent decades, it seems likely that further ones still remain to be discovered.

The same technologies that will pose these risks will also help us to mitigate some risks. Biotechnology can help us develop better diagnostics, vaccines, and anti-viral drugs. Molecular nanotechnology could offer even stronger prophylactics.[28] Superintelligent machines may be the last invention that human beings ever need to make, since a superintelligence, by definition, would be far more effective than a human brain in practically all intellectual endeavors, including strategic thinking, scientific analysis, and technological creativity.[29] In addition to creating and mitigating risks, these powerful technological capabilities would also affect the human condition in many other ways.

Extinction risks constitute an especially severe subset of what could go badly wrong for humanity. There are many possible global catastrophes that would cause immense worldwide damage, maybe even the collapse of modern civilization, yet fall short of terminating the human species. An all-out nuclear war between Russia and the United States might be an example of a global catastrophe that would be unlikely to result in extinction. A terrible pandemic with high virulence and 100% mortality rate among infected individuals might be another example: if some groups of humans could successfully quarantine themselves before being exposed, human extinction could be avoided even if, say, 95% or more of the world's population succumbed. What distinguishes extinction and other existential catastrophes is that a comeback is impossible. A non-existential disaster causing the breakdown of global civilization is, from the perspective of humanity as a whole, a potentially recoverable setback: a giant massacre for man, a small misstep for mankind.

An existential catastrophe is therefore qualitatively distinct from a "mere" collapse of global civilization, although in terms of our moral and prudential

26 (Drexler 1985). Drexler is even more concerned about the potential misuse of tools based on advanced nanotechnology to control and oppress populations than he is about the possibility that nanotechnology weapons systems would be used to directly cause human extinction; (Drexler 2007), p. 57.

27 (Bostrom 2002b; Yudkowsky 2007)

28 (Freitas 1999)

29 (Bostrom 1998)

attitudes perhaps we should simply view both as unimaginably bad outcomes.[30] One way that civilization collapse could be a significant feature in the larger picture for humanity, however, is if it formed part of a repeating pattern. This takes us to the second family of scenarios: recurrent collapse.

## Recurrent collapse

Environmental threats seem to have displaced nuclear holocaust as the chief specter haunting the public imagination. Current-day pessimists about the future often focus on the environmental problems facing the growing world population, worrying that our wasteful and polluting ways are unsustainable and potentially ruinous to human civilization. The credit for having handed the environmental movement its initial impetus is often given to Rachel Carson, whose book *Silent Spring* (1962) sounded the alarm on pesticides and synthetic chemicals that were being released into the environment with allegedly devastating effects on wildlife and human health.[31] The environmentalist forebodings swelled over the decade. Paul Ehrlich's book *Population Bomb*, and the Club of Rome report *Limits to Growth*, which sold 30 million copies, predicted economic collapse and mass starvation by the eighties or nineties as the results of population growth and resource depletion.[32]

In recent years, the spotlight of environmental concern has shifted to global climate change. Carbon dioxide and other greenhouse gases are accumulating in the atmosphere, where they are expected to cause a warming of Earth's climate and a concomitant rise in sea water levels. The more recent report by the United Nations' Intergovernmental Panel on Climate Change, which represents the most authoritative assessment of current scientific opinion, attempts to estimate the increase in global mean temperature that would be expected by the end of this century under the assumption that no efforts at mitigation are made. The final estimate is fraught with uncertainty because of uncertainty about what the default rate of emissions of greenhouse gases will be over the century, uncertainty about the climate sensitivity parameter, and uncertainty about other factors. The IPCC therefore expresses its assessment in terms of six different climate scenarios based on different models and different assumptions. The "low" model predicts a mean global warming of +1.8°C (uncertainty range 1.1°C to 2.9°C); the "high" model predicts warming by +4.0°C (2.4°C to 6.4°C).[33] Estimated sea level rise predicted by these two most extreme scenarios among the six considered is 18 to 38 cm, and 26 to 59 cm, respectively.[34]

30 How much worse would an existential risk be than an event that merely killed 99% of all humans but allowed for eventual recovery? The answer requires a theory of value. See e.g. (Parfit 1984; Bostrom 2003a, 2007a).

31 (Carson 1962)

32 (Ehrlich 1968; Meadows and Club of Rome. 1972)

33 (Solomon et al. 2007), p. 749

34 Ibid., p. 750

While this prognosis might well justify a range of mitigation policies, it is important to maintain a sense of perspective when we are considering the issue from a "future of humanity" point of view. Even the *Stern Review on the Economics of Climate Change*, a report prepared for the British Government which has been criticized by some as overly pessimistic, estimates that under the assumption of business-as-usual with regard to emissions, global warming will reduce welfare by an amount equivalent to a permanent reduction in per capita consumption of between 5 and 20%.[35] In absolute terms, this would be a huge harm. Yet over the course of the twentieth century, world GDP grew by some 3,700%, and per capita world GDP rose by some 860%.[36] It seems safe to say that (absent a radical overhaul of our best current scientific models of the Earth's climate system) whatever negative economic effects global warming will have, they will be completely swamped by other factors that will influence economic growth rates in this century.

There have been a number of attempts by scholars to explain societal collapse – either as a case study of some particular society, such as Gibbon's classic *Decline and Fall of the Roman Empire* – or else as an attempt to discover failure modes applying more generally.[37] Two examples of the latter genre include Joseph Tainter's *Collapse of Complex Societies*, and Jared Diamond's more recent *Collapse: How Societies Choose to Fail or Succeed*. Tainter notes that societies need to secure certain resources such as food, energy, and natural resources in order to sustain their populations.[38] In their attempts to solve this supply problem, societies may grow in complexity – for example, in the form of bureaucracy, infrastructure, social class distinction, military operations, and colonies. At some point, Tainter argues, the marginal returns on these investments in social complexity become unfavorable, and societies that do not manage to scale back when their organizational overheads become too large eventually face collapse.

Diamond argues that many past cases of societal collapse have involved environmental factors such as deforestation and habitat destruction, soil problems, water management problems, overhunting and overfishing, the effects of introduced species, human population growth, and increased per-capita impact of people.[39] He also suggests four new factors that may contribute to the collapse of present and future societies: human-caused climate change, but also build-up of toxic chemicals in the environment, energy shortages, and the full utilization of the Earth's photosynthetic capacity. Diamond draws attention to the danger of "creeping normalcy", referring to the phenomenon

35 (Stern and Great Britain Treasury 2006); for references to critiques thereof, see e.g. (Nordhaus 2007; Cox and Vadon 2007).

36 These numbers, which are of course approximate, are calculated from data presented in (De Long and Olney 2006); see also (De Long 1998).

37 (Gibbon and Kitchin 1777)

38 (Tainter 1988)

39 (Diamond 2005)

of a slow trend being concealed within noisy fluctuations, so that a detrimental outcome that occurs in small, almost unnoticeable steps may be accepted or come about without resistance even if the same outcome, had it come about in one sudden leap, would have evoked a vigorous response.[40]

We need to distinguish different classes of scenarios involving societal collapse. First, we may have a merely local collapse: individual societies can collapse, but this is unlikely to have a determining effect on the future of humanity if other advanced societies survive and take up where the failed societies left off. All historical examples of collapse have been of this kind. Second, we might suppose that new kinds of threat (e.g. nuclear holocaust or catastrophic changes in the global environment) or the trend towards globalization and increased interdependence of different parts of the world create a vulnerability to human civilization as a whole. Suppose that a global societal collapse were to occur. What happens next? If the collapse is of such a nature that a new advanced global civilization can *never* be rebuilt, the outcome would qualify as an existential disaster. However, it is hard to think of a plausible collapse which the human species survives but which nevertheless makes it permanently impossible to rebuild civilization. Supposing, therefore, that a new technologically advanced civilization is eventually rebuilt, what is the fate of this resurgent civilization? Again, there are two possibilities. The new civilization might avoid collapse; and in the following two sections we will examine what could happen to such a sustainable global civilization. Alternatively, the new civilization collapses again, and the cycle repeats. If eventually a sustainable civilization arises, we reach the kind of scenario that the following sections will discuss. If instead one of the collapses leads to extinction, then we have the kind of scenario that was discussed in the previous section. The remaining case is that we face a cycle of indefinitely repeating collapse and regeneration (see figure 1).

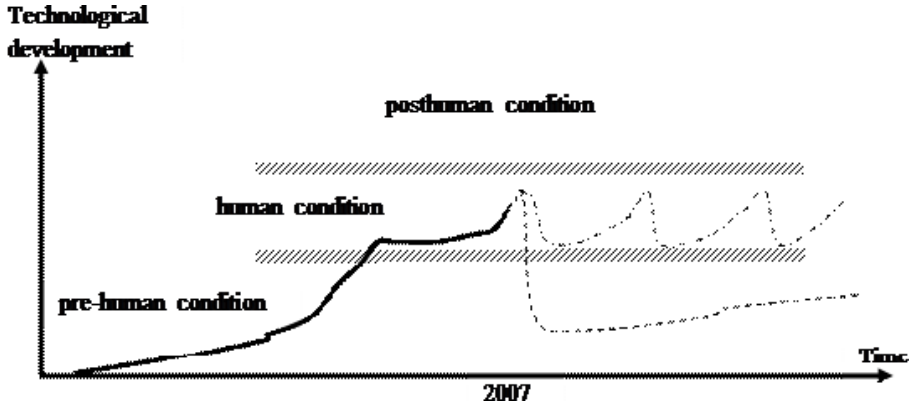



**Figure 1.** Schematic of two types of scenario for the future of humanity. One line illustrates an annihilation scenario in which the human species is destroyed

40 Ibid., p. 425.

a short while (perhaps a few decades) after the present time. The other line illustrates a recurrent collapse scenario, in which human civilization oscillates indefinitely within the range of technological development characteristic of a human condition. (The y-axis is not an index of value; "up" is not necessarily better.)

While there are many conceivable explanations for why an advanced society might collapse, only a subset of these explanations could plausibly account for an unending pattern of collapse and regeneration. An explanation for such a cycle could not rely on some contingent factor that would apply to only some advanced civilizations and not others, or to a factor that an advanced civilization would have a realistic chance of counteracting; for if such a factor were responsible, one would expect that the collapse-regeneration pattern would at some point be broken when the right circumstances finally enabled an advanced civilization to overcome the obstacles to sustainability. Yet at the same time, the postulated cause for collapse could not be so powerful as to cause the extinction of the human species.

A recurrent collapse scenario consequently requires a carefully calibrated homeostatic mechanism that keeps the level of civilization confined within a relatively narrow interval, as illustrated in figure 1. Even if humanity were to spend many millennia on such an oscillating trajectory, one might expect that eventually this phase would end, resulting in either the permanent destruction of humankind, or the rise of a stable sustainable global civilization, or the transformation of the human condition into a new "posthuman" condition. We turn now to the second of these possibilities, that the human condition will reach a kind of stasis, either immediately or after undergoing one or more cycles of collapse-regeneration.

## Plateau

Figure 2 depicts two possible trajectories, one representing an increase followed by a permanent plateau, the other representing stasis at (or close to) the current status quo.

The static view is implausible. It would imply that we have recently arrived at the final human condition even at a time when change is exceptionally rapid: "What we do know," writes distinguished historian of technology Vaclav Smil, "is that the past six generations have amounted to the most rapid and the most profound change our species has experienced in its 5,000 years of recorded history."[41] The static view would also imply a radical break with several long-established trends. If the world economy continues to grow at the same pace as in the last half century, then by 2050 the world will be seven times richer than it is today. World population is predicted to increase to just over 9 billion in 2050, so average wealth would also increase dramatically.[42] Extrapolating

41 (Smil 2006), p. 311.

42 (United Nations Population Division 2004)

further, by 2100 the world would be almost 50 times richer than today. A single modest-sized country might then have as much wealth as the entire world has at the present. Over the course of human history, the doubling time of the world economy has been drastically reduced on several occasions, such as in the agricultural transition and the Industrial Revolution. Should another such transition occur in this century, the world economy might be several orders of magnitude larger by the end of the century.[43]

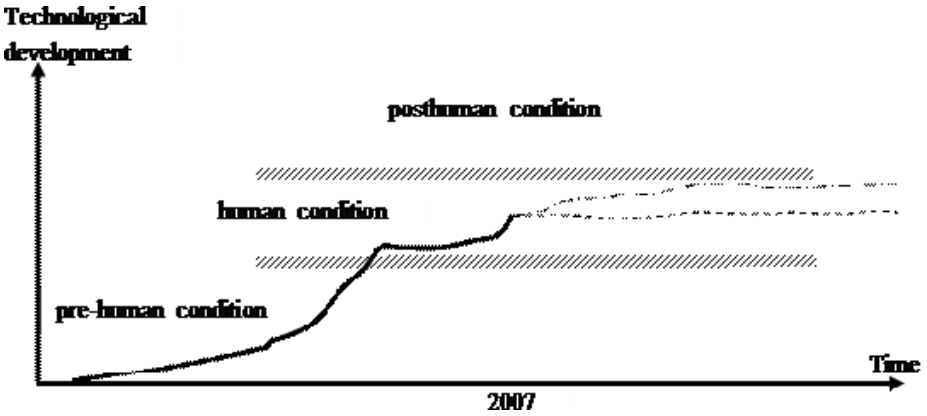



**Figure 2.** Two trajectories: increase followed by plateau; or stasis at close to the current level.

Another reason for assigning a low probability to the static view is that we can foresee various specific technological advances that will give humans important new capacities. Virtual reality environments will constitute an expanding fraction of our experience. The capability of recording, surveillance, biometrics, and data mining technologies will grow, making it increasingly feasible to keep track of where people go, whom they meet, what they do, and what goes on inside their bodies.[44]

Among the most important potential developments are ones that would enable us to alter our biology directly through technological means.[45] Such interventions could affect us more profoundly than modification of beliefs, habits, culture, and education. If we learn to control the biochemical processes of human senescence, healthy lifespan could be radically prolonged. A person with the age-specific mortality of a 20-year-old would have a life expectancy of about a thousand years. The ancient but hitherto mostly futile quest for happiness could meet with success if scientists could develop safe and effective methods of controlling the brain circuitry responsible for subjective well-being.[46] Drugs and other neurotechnologies could make it increasingly

43 (Hanson 2000)

44 (Brin 1998)

45 (Bostrom 2005, 2007c)

46 (Pearce 2004)

feasible for users to shape themselves into the kind of people they want to be by adjusting their personality, emotional character, mental energy, romantic attachments, and moral character.[47] Cognitive enhancements might deepen our intellectual lives.[48]

Nanotechnology will have wide-ranging consequences for manufacturing, medicine, and computing.[49] Machine intelligence, to be discussed further in the next section, is another potential revolutionary technology. Institutional innovations such as prediction markets might improve the capability of human groups to forecast future developments, and other technological or institutional developments might lead to new ways for humans to organize more effectively.[50] The impacts of these and other technological developments on the character of human lives are difficult to predict, but that they will have such impacts seems a safe bet.

Those who believe that developments such as those listed will not occur should consider whether their skepticism is really about ultimate feasibility or merely about timescales. Some of these technologies will be difficult to develop. Does that give us reason to think that they will never be developed? Not even in 50 years? 200 years? 10,000 years? Looking back, developments such as language, agriculture, and perhaps the Industrial Revolution may be said to have significantly changed the human condition. There are at least a thousand times more of us now; and with current world average life expectancy at 67 years, we live perhaps three times longer than our Pleistocene ancestors. The mental life of human beings has been transformed by developments such as language, literacy, urbanization, division of labor, industrialization, science, communications, transport, and media technology.

The other trajectory in figure 2 represents scenarios in which technological capability continues to grow significantly beyond the current level before leveling off below the level at which a fundamental alteration of the human condition would occur. This trajectory avoids the implausibility of postulating that we have just now reached a permanent plateau of technological development. Nevertheless, it does propose that a permanent plateau will be reached not radically far above the current level. We must ask what could cause technological development to level off at that stage.

One conceptual possibility is that development beyond this level is impossible because of limitations imposed by fundamental natural laws. It appears, however, that the physical laws of our universe permit forms of organization

47 (Pearce 2004)

48 (Bostrom and Ord 2006; Bostrom and Sandberg 2007)

49 Molecular nanotechnology (aka molecular manufacturing, or machine-phase nanotechnology) is one area where a considerable amount of "theoretical applied science" has been done, although this has not yet resulted in a consensus about the feasibility of this anticipated technology; see e.g. (Drexler 1992).

50 (Hanson 1995; Wolfers and Zitzewitz 2004)

that would qualify as a posthuman condition (to be discussed further in the next section). Moreover, there appears to be no fundamental obstacle to the development of technologies that would make it possible to build such forms of organization.[51] Physical impossibility, therefore, is not a plausible explanation for why we should end up on either of the trajectories depicted in figure 2.

Another potential explanation is that while theoretically possible, a posthuman condition is just too difficult to attain for humanity ever to be able to get there. For this explanation to work, the difficulty would have to be of a certain kind. If the difficulty consisted merely of there being a large number of technologically challenging steps that would be required to reach the destination, then the argument would at best suggest that it will take a long time to get there, not that we never will. Provided the challenge can be divided into a sequence of individually feasible steps, it would seem that humanity could eventually solve the challenge given enough time. Since at this point we are not so concerned with timescales, it does not appear that technological difficulty of this kind would make any of the trajectories in figure 2 a plausible scenario for the future of humanity.

In order for technological difficulty to account for one of the trajectories in figure 2, the difficulty would have to be of a sort that is not reducible to a long sequence of individually feasible steps. If all the pathways to a posthuman condition required technological capabilities that could be attained only by building enormously complex, error-intolerant systems of a kind which could not be created by trial-and-error or by assembling components that could be separately tested and debugged, then the technological difficulty argument would have legs to stand on. Charles Perrow argued in *Normal Accidents* that efforts to make complex systems safer often backfire because the added safety mechanisms bring with them additional complexity which creates additional opportunities for things to go wrong when parts and processes interact in unexpected ways.[52] For example, increasing the number of security personnel on a site can increase the "insider threat", the risk that at least one person on the inside can be recruited by would-be attackers.[53] Along similar lines, Jaron Lanier has argued that software development has run into a kind of complexity barrier.[54] An informal argument of this kind has also been made against the feasibility of molecular manufacturing.[55]

Each of these arguments about complexity barriers is problematic. And in order to have an explanation for why humanity's technological development should level off before a posthuman condition is reached, it is not sufficient to show that *some* technologies run into insuperable complexity barriers. Rather,

51 See e.g. (Bostrom 2003b; Moravec 1999; Drexler 1985; Kurzweil 2005)

52 (Perrow 1984)

53 See e.g. (Sagan 2004).

54 (Lanier 2000)

55 (Burkhead 1999)

it would have to be shown that *all* technologies that would enable a posthuman condition (biotechnology, nanotechnology, artificial intelligence, etc.) will be blocked by such barriers. That seems an unlikely proposition. Alternatively, one might try to build an argument based on complexity barriers for social organization in general rather than for particular technologies – perhaps something akin to Tainter's explanation of past cases of societal collapse, mentioned in the previous section. In order to produce the trajectories in figure 2, however, the explanation would have to be modified to allow for stagnation and plateauing rather than collapse. One problem with this hypothesis is that it is unclear that the development of the technologies requisite to reach a posthuman condition would necessarily require a significant increase in the complexity of social organization beyond its present level.

A third possible explanation is that even if a posthuman condition is both theoretically possible and practically feasible, humanity might "decide" not to pursue technological development beyond a certain level. One could imagine systems, institutions, or attitudes emerging which would have the effect of blocking further development, whether by design or as an unintended consequence. Yet an explanation rooted in unwillingness for technological advancement would have to overcome several challenges. First, how does enough unwillingness arise to overcome what at the present appears like an inexorable process of technological innovation and scientific research? Second, how does a decision to relinquish development get implemented globally in a way that leaves no country and no underground movement able to continue technological research? Third, how does the policy of relinquishment avoid being overturned, even on timescales extending over tens of thousands of years and beyond? Relinquishment would have to be global and permanent in order to account for a trajectory like one of those represented in figure 2. A fourth difficulty emerges out of the three already mentioned: the explanation for how the aversion to technological advancement arises, how it gets universally implemented, and how it attains permanence, would have to avoid postulating causes that in themselves would usher in a posthuman condition. For example, if the explanation postulated that powerful new mind-control technologies would be deployed globally to change people's motivation, or that an intensive global surveillance system would be put in place and used to manipulate the direction of human development along a predetermined path, one would have to wonder whether these interventions, or their knock-on effects on society, culture, and politics, would not themselves alter the human condition in sufficiently fundamental ways that the resulting condition would qualify as posthuman.

To argue that stasis and plateau are relatively unlikely scenarios is not inconsistent with maintaining that *some aspects* of the human condition will remain unchanged. For example, Francis Fukuyama argued in *The End of History and the Last Man* that the endpoint of mankind's ideological

evolution has essentially been reached with the end of the Cold War.[56] Fukuyama suggested that Western liberal democracy is the final form of human government, and that while it would take some time for this ideology to become completely universalized, secular free-market democracy will in the long term become more and more prevalent. In his more recent book *Our Posthuman Future*, he adds an important qualification to his earlier thesis, namely that direct technological modification of human nature could undermine the foundations of liberal democracy.[57] But be that as it may, the thesis that liberal democracy (or any other political structure) is the final form of government is consistent with the thesis that the general condition for intelligent Earth-originating life will not remain a *human* condition for the indefinite future.

## Posthumanity

An explication of what has been referred to as "posthuman condition" is overdue. In this paper, the term is used to refer to a condition which has at least one of the following characteristics:

- Population greater than 1 trillion persons
- Life expectancy greater than 500 years
- Large fraction of the population has cognitive capacities more than two standard deviations above the current human maximum
- Near-complete control over the sensory input, for the majority of people for most of the time
- Human psychological suffering becoming a rare occurrence
- Any change of magnitude or profundity comparable to that of one of the above

This definition's vagueness and arbitrariness may perhaps be excused on grounds that the rest of this paper is at least equally schematic. In contrast to some other explications of "posthumanity", the one above does not require direct modification of human nature.[58] This is because the relevant concept for the present discussion is that of a level of technological or economic development that would involve a radical change in the human condition, whether the change was wrought by biological enhancement or other causes.

56 (Fukuyama 1992)

57 (Fukuyama 2002)

58 E.g. (Bostrom 2003b, 2007c)

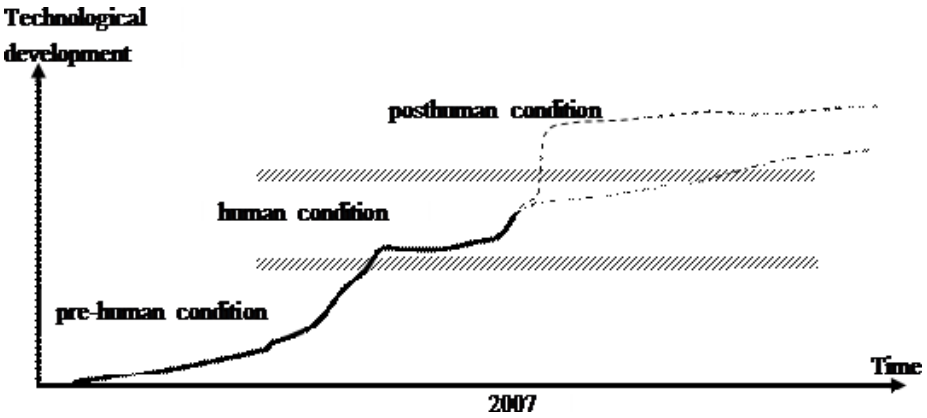



**Figure 3.** A singularity scenario, and a more incremental ascent into a posthuman condition.

The two dashed lines in figure 3 differ in steepness. One of them depicts slow gradual growth that in the fullness of time rises into the posthuman level and beyond. The other depicts a period of extremely rapid growth in which humanity abruptly transitions into a posthuman condition. This latter possibility can be referred to as *the singularity hypothesis*.[59] Proponents of the singularity hypothesis usually believe not only that a period of extremely rapid technological development will usher in posthumanity suddenly, but also that this transition will take place soon – within a few decades. Logically, these two contentions are quite distinct.

In 1958, Stanislaw Ulam, a Polish-born American mathematician, referring to a meeting with John von Neumann, wrote:

> One conversation centered on the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.[60]

The idea of a technological singularity tied specifically to artificial intelligence was perhaps first clearly articulated by the statistician I. J. Good in 1965:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and

59 "Singularity" is to be interpreted here not in its strict mathematical meaning but as suggesting extreme abruptness. There is no claim that any of the quantities involved would become literally infinite or undefined.

60 (Ulam 1958)

> the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the *last* invention that man need ever make… It is more probable than not that, within the twentieth century, an ultraintelligent machine will be built…[61]

Mathematician and science fiction writer Vernor Vinge elaborated on this idea in his 1993 essay *The Coming Technological Singularity*, adjusting the timing of Good's prediction:

> Within thirty years, we will have the technological means to create superhuman intelligence. Shortly thereafter, the human era will be ended.[62]

Vinge considered several possible avenues to superintelligence, including AI in individual machines or computer networks, computer/human interfaces, and biological improvement of the natural human intellect. An important part of both Good's and Vinge's reasoning is the idea of a strong positive feedback-loop as increases in intelligence lead to increased ability to make additional progress in intelligence-increasing technologies. ("Intelligence" could here be understood as a general rubric for all those mental faculties that are relevant for developing new technologies, thus including for example creativity, work capacity, and the ability to write a persuasive case for funding.)

Skeptics of the singularity hypothesis can object that while *ceteris paribus* greater intelligence would lead to faster technological progress, there is an additional factor at play which may slow things down, namely that the easiest improvements will be made first, and that after the low-hanging fruits have all been picked, each subsequent improvement will be more difficult and require a greater amount of intellectual capability and labor to achieve. The mere existence of positive feedback, therefore, is not sufficient to establish that an intelligence explosion would occur once intelligence reaches some critical magnitude.

To assess the singularity hypothesis one must consider more carefully what kinds of intelligence-increasing interventions might be feasible and how closely stacked these interventions are in terms of their difficulty. Only if intelligence growth could exceed the growth in difficulty level for each subsequent improvement could there be a singularity. The period of rapid intelligence growth would also have to last long enough to usher in a posthuman era before running out of steam.

It might be easiest to assess the prospect for an intelligence explosion if we focus on the possibility of quantitative rather than qualitative improvements in intelligence. One interesting pathway to greater intelligence illustrating such quantitative growth – and one that Vinge did not discuss – is uploading.

61 (Good 1965)

62 (Vinge 1993)

Uploading refers to the use of technology to transfer a human mind to a computer. This would involve the following steps: First, create a sufficiently detailed scan of a particular human brain, perhaps by feeding vitrified brain tissue into an array of powerful microscopes for automatic slicing and scanning. Second, from this scanning data, use automatic image processing to reconstruct the 3-dimensional neuronal network that implemented cognition in the original brain, and combine this map with neurocomputational models of the different types of neurons contained in the network. Third, emulate the whole computational structure on a powerful supercomputer (or cluster). If successful, the procedure would result in a qualitative reproduction of the original mind, with memory and personality intact, onto a computer where it would now exist as software.[63] This mind could either inhabit a robotic body or live in virtual reality. In determining the prerequisites for uploading, a tradeoff exists between the power of the scanning and simulation technology, on the one hand, and the degree of neuroscience insight on the other. The worse the resolution of the scan, and the lower the computing power available to simulate functionally possibly irrelevant features, the more scientific insight would be needed to make the procedure work. Conversely, with sufficiently advanced scanning technology and enough computing power, it might be possible to brute-force an upload even with fairly limited understanding of how the brain works – perhaps a level of understanding representing merely an incremental advance over the current state of the art.

One obvious consequence of uploading is that many copies could be created of one uploaded mind. The limiting resource is computing power to store and run the upload minds. If enough computing hardware already exists or could rapidly be built, the upload population could undergo explosive growth: the replication time of an upload need be no longer than the time it takes to make a copy of a big piece of software, perhaps minutes or hours – a vast speed-up compared to biological human replication. And the upload replica would be an exact copy, possessing from birth all the skills and knowledge of the original. This could result in rapidly exponential growth in the supply of highly skilled labor.[64] Additional acceleration is likely to result from improvements in the computational efficiency of the algorithms used to run the uploaded minds. Such improvements would make it possible to create faster-thinking uploads, running perhaps at speeds thousands or millions of times that of an organic brain.

If uploading is technologically feasible, therefore, a singularity scenario involving an intelligence explosion and very rapid change seems realistic based only on

63 I use the term "qualitative reproduction" advisedly, in order to sidestep the philosophical questions of whether the original mind could be quantitatively the same mind as the upload, and whether the uploaded person could survive the procedure and continue to live as an upload. The relevance of uploading to the present argument does not depend on the answers to these questions.

64 (Hanson 1994). Absent regulation, this would lead to a precipitous drop in wages.

the possibility of quantitative growth in machine intelligence.[65] The harder-to-evaluate prospect of qualitative improvements adds some further credence to the singularity hypothesis.[66]

Uploading would almost certainly produce a condition that would qualify as "posthuman" in this paper's terminology, for example on grounds of population size, control of sensory input, and life expectancy. (A human upload could have an indefinitely long lifespan as it would not be subject to biological senescence, and periodic backup copies could be created for additional security.) Further changes would likely follow swiftly from the productivity growth brought about by the population expansion. These further changes may include qualitative improvements in the intelligence of uploads, other machine intelligences, and remaining biological human beings.[67]

Inventor and futurist Ray Kurzweil has argued for the singularity hypothesis on somewhat different grounds. His most recent book, *The Singularity Is Near*, is an update of his earlier writings.[68] It covers a vast range of ancillary topics related to radical future technological prospects, but its central theme is an attempt to demonstrate "the law of accelerating returns", which manifests itself as exponential technological progress. Kurzweil plots progress in a variety of areas, including computing, communications, and biotechnology, and in each case finds a pattern similar to Moore's law for microchips: performance grows as an exponential with a short doubling time (typically a couple of years). Extrapolating these trend lines, Kurzweil infers that a technological singularity is due around the year 2045.[69] While machine intelligence features as a prominent factor in Kurzweil's forecast, his singularity scenario differs from that of Vinge in being more gradual: not a virtually-overnight total transformation resulting from runaway self-improving artificial intelligence, but a steadily accelerating pace of general technological advancement.

Several critiques could be leveled against Kurzweil's reasoning. First, one might of course doubt that present exponential trends will continue for another four

65 The antecedent of the conditional ("if uploading is technologically feasible –") includes, of course, assumptions of a metaphysical nature, such as the assumption that a computer could in principle manifest the same level of intelligence as a biological human brain. However, in order to see that uploading would have wide-ranging practical ramifications, it is not necessary to assume that uploads would have qualia or subjective conscious experiences. The question of upload qualia would be important, though, in assessing the meaning and value of scenarios in which a significant percentage of the population of intelligent beings are machine-based.

66 To say something more definite about the probability of a singularity, we would at this stage of the analysis have to settle on a more unambiguous definition of the term.

67 The distinction between quantitative and qualitative improvements may blur in this context. When I suggest that qualitative changes might occur, I am not referring to a strict mathematical concept like Turing computability, but to a looser idea of an improvement in intelligence that is not aptly characterized as a mere speed-up.

68 (Kurzweil 2005)

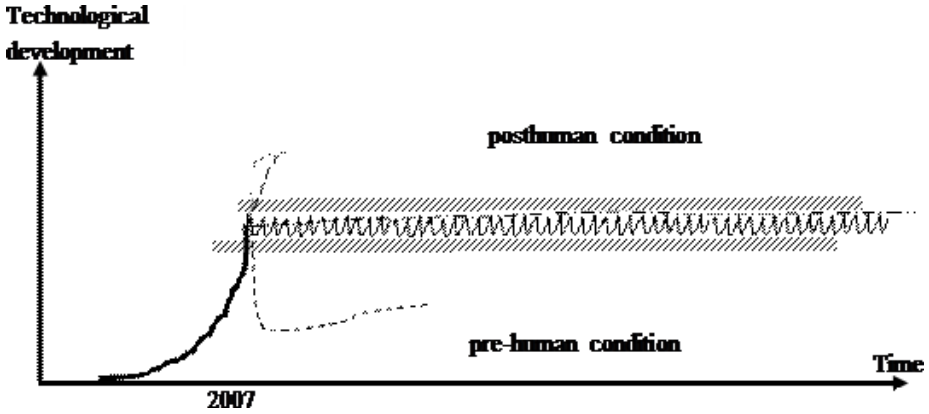
69 Note that the expected arrival time of the singularity has receded at a rate of roughly one year per year. Good, writing in 1965, expected it before 2000. Vinge, writing in 1993, expected it before 2023. Kurzweil, writing in 2005, expects it by 2045.

decades. Second, while it is possible to identify certain fast-growing areas, such as IT and biotech, there are many other technology areas where progress is much slower. One could argue that to get an index of the overall pace of technological development, we should look not at a hand-picked portfolio of hot technologies, but instead at economic growth, which implicitly incorporates all productivity-enhancing technological innovations, weighted by their economic significance. In fact, the world economy has also been growing at a roughly exponential rate since the Industrial Revolution; but the doubling time is much longer, approximately 20 years.[70] Third, if technological progress is exponential, then the current rate of technological progress must be vastly greater than it was in the remote past. But it is far from clear that this is so. Vaclav Smil – the historian of technology who, as we saw, has argued that the past six generations have seen the most rapid and profound change in recorded history – maintains that the 1880s was the most innovative decade of human history.[71]

## The longer term

The four families of scenarios we have considered – extinction, recurrent collapse, plateau, and posthumanity – could be modulated by varying the timescale over which they are hypothesized to occur. A few hundred years or a few thousand years might already be ample time for the scenarios to have an opportunity to play themselves out. Yet such an interval is a blip compared to the lifetime of the universe. Let us therefore zoom out and consider the longer term prospects for humanity.

The first thing to notice is that the longer the time scale we are considering, the less likely it is that technological civilization will remain within the zone we termed "the human condition" throughout. We can illustrate this point graphically by redrawing the earlier diagrams using an expanded scale on the two axes (figure 4).




70 (De Long 1998)

71 (Smil 2006), p. 131

**Figure 4.** The scenarios presented in previous figures are here represented with a time axis that is slightly closer to linear and a y-axis that slightly better reveals how narrow a band the "human condition" is among all the possible levels of organismic and technological development. The graph is still a mere schematic, not a strictly quantitative representation. Note how the scenarios that postulate that the human condition will continue to hold indefinitely begin to look increasingly peculiar as we adjust the scales to reveal more of the larger picture.

The extinction scenario is perhaps the one least affected by extending the timeframe of consideration. If humanity goes extinct, it stays extinct.[72] The cumulative probability of extinction increases monotonically over time. One might argue, however, that the current century, or the next few centuries, will be a critical phase for humanity, such that if we make it through this period then the life expectancy of human civilization could become extremely high. Several possible lines of argument would support this view. For example, one might believe that superintelligence will be developed within a few centuries, and that, while the creation of superintelligence will pose grave risks, once that creation and its immediate aftermath have been survived, the new civilization would have vastly improved survival prospects since it would be guided by superintelligent foresight and planning. Furthermore, one might believe that self-sustaining space colonies may have been established within such a timeframe, and that once a human or posthuman civilization becomes dispersed over multiple planets and solar systems, the risk of extinction declines. One might also believe that many of the possible revolutionary technologies (not only superintelligence) that *can* be developed *will* be developed within the next several hundred years; and that if these technological revolutions are destined to cause existential disaster, they would already have done so by then.

The recurrent collapse scenario becomes increasingly unlikely the longer the timescale, for reasons that are apparent from figure 4. The scenario postulates that technological civilization will oscillate continuously within a relatively narrow band of development. If there is any chance that a cycle will either break through to the posthuman level or plummet into extinction, then there is for each period a chance that the oscillation will end. Unless the chance of such a breakout converges to zero at a sufficiently rapid rate, then with probability one the pattern will *eventually* be broken. At that point the pattern might degenerate into one of the other ones we have considered.

The plateau scenarios are similar to the recurrent collapse scenario in that the level of civilization is hypothesized to remain confined within a narrow range; and the longer the timeframe considered, the smaller the probability that the level of technological development will remain within this range. But compared to the recurrent collapse pattern, the plateau pattern might be thought to have

72 It is possible that if humanity goes extinct, another intelligent species might evolve on Earth to fill the vacancy. The fate of such a possible future substitute species, however, would not strictly be part of the future of *humanity*.

a bit more staying power. The reason is that the plateau pattern is consistent with a situation of complete stasis – such as might result, for example, from the rise of a very stable political system, propped up by greatly increased powers of surveillance and population control, and which for one reason or another opts to preserve its status quo. Such stability is inconsistent with the recurrent collapse scenario.

The cumulative probability of posthumanity, like that of extinction, increases monotonically over time. By contrast to extinction scenarios, however, there is a possibility that a civilization that has attained a posthuman condition will later revert to a human condition. For reasons paralleling those suggested earlier for the idea that the annual risk of extinction will decline substantially after certain critical technologies have been developed and after self-sustaining space colonies have been created, one might maintain that the annual probability that a posthuman condition would revert to a human condition will likewise decline over time.[73]

## References


- Bostrom, N. (1998) "How Long Before Superintelligence?" *International Journal of Futures Studies* 2.
- — (2002a) *Anthropic Bias: Observation Selection Effects in Science and Philosophy* (New York: Routledge).
- — (2002b) "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", *Journal of Evolution and Technology* 9.
- — (2002c) "Self-Locating Belief in Big Worlds: Cosmology's Missing Link to Observation", *Journal of Philosophy* 99 (12):607–623.
- — (2003a) "Astronomical Waste: The Opportunity Cost of Delayed Technological Development", *Utilitas* 15 (3):308-314.
- — *The Transhumanist FAQ: v 2.1.* World Transhumanist Association 2003b. Available from https://nickbostrom.com/views/transhumanist.pdf.
- — (2005) "Transhumanist Values", *Review of Contemporary Philosophy* 4 (1-2):87-101.
- — (2006) "Quantity of Experience: Brain-Duplication and Degrees of Consciousness", *Minds and Machines* 16 (2):185-200.
- — (2007a) "Infinite Ethics", in, *Working manuscript.* Available from https://nickbostrom.com/ethics/infinite.pdf.


73 I am grateful to Rebecca Roache for research assistance and to her and Nick Shackel for helpful comments on an earlier draft.

- — (2007b) "Technological Revolutions: Ethics and Policy in the Dark", in Nigel M. de S. Cameron (ed.), *Nanotechnology and Society* (John Wiley).
- — (2007c) "Why I Want to be a Posthuman When I Grow Up", in Bert Gordijn and Ruth Chadwick (eds.), *Medical Enhancement and Posthumanity* (Springer).
- Bostrom, N., and Ord, T. (2006) "The Reversal Test: Eliminating Status Quo Bias in Bioethics", *Ethics* 116 (4):656-680.
- Bostrom, N., and Sandberg, A. (2007) "Cognitive Enhancement: Methods, Ethics, Regulatory Challenges", *Science and Engineering Ethics* forthcoming.
- Brin, D. (1998) *The Transparent Society* (Reading, Mass.: Addison-Wesley).
- Bureau, U. S. C. *Historical Estimates of World Population* 2007. Available from https://www.census.gov/data/tables/time-series/demo/international-programs/historical-est-worldpop.html.
- Burkhead, L. *Nanotechnology without Genies* 1999. Available from http://www.geniebusters.org/00_contents.htm.
- Carson, R. (1962) *Silent spring* (Boston: Houghton Mifflin).
- Cox, S., and Vadon, R. (2007) "Running the rule over Stern's numbers", in, *BBC Radio 4, The Investigation*. Available from http://news.bbc.co.uk/1/hi/sci/tech/6295021.stm.
- Crow, M. M., and Sarewitz, D. (2001) "Nanotechnology and Societal Transformation", in Albert H. Teich, Stephen D. Nelson, Celia McEnaney and Stephen J. Lita (eds.), *AAAS Science and Technology Policy Yearbook* (Washington, DC: American Association for the Advancement of Science), 89-101.
- De Long, J. B. (1998) "Estimating World GDP, One Million B.C. - Present", in, *Electronic document*. Available from http://econ161.berkeley.edu/TCEH/1998_Draft/World_GDP/Estimating_World_GDP.html.
- De Long, J. B., and Olney, M. L. (2006) *Macroeconomics*. 2nd ed (Boston: McGraw-Hill).
- Diamond, J. M. (2005) *Collapse: how societies choose to fail or succeed* (New York: Viking).
- Drexler, E. (1992) *Nanosystems: Molecular Machinery, Manufacturing, and Computation* (New York: John Wiley & Sons, Inc.).
- — (2003) "Nanotechnology Essays: Revolutionizing the Future of Technology (Revised 2006)", *AAAS EurekAlert! InContext* April.

- — (2007) "The stealth threat: an interview with K. Eric Drexler", *Bulletin of the Atomic Scientists* 68 (1):55-58.
- Drexler, K. E. (1985) *Engines of Creation: The Coming Era of Nanotechnology* (London: Fourth Estate).
- Ehrlich, P. R. (1968) *The population bomb* (New York: Ballantine Books).
- Freitas, R. A. (1999) *Nanomedicine* (Austin, TX: Landes Bioscience).
- Fukuyama, F. (1992) *The end of history and the last man* (New York: Free Press).
- — (2002) *Our Posthuman Future: Consequences of the Biotechnology Revolution* (Farrar, Straus and Giroux).
- Gibbon, E., and Kitchin, T. (1777) *The history of the decline and fall of the Roman empire: in twelve volumes*. A new edition ed. 12 vols (London: Printed for Lackington, Allen, and Co.).
- Good, I. J. (1965) "Speculations Concerning the First Ultraintelligent Machine", *Advances in Computers* 6:31-88.
- Hanson, R. (1994) "What If Uploads Come First: The Crack of a Future Dawn", *Extropy* 6 (2).
- — (1995) "Could Gambling Save Science? Encouraging an Honest Consensus", *Social Epistemology* 9:1:3-33.
- — (2000) "Long-term growth as a sequence of exponential modes", *Working manuscript*.
- Heilbroner, R. L. (1995) *Visions of the future: the distant past, yesterday, today, tomorrow* (New York: Oxford University Press).
- Hughes, J. (2007) "Millennial Tendencies in Responses to Apocalyptic Threats", in Nick Bostrom and Milan Cirkovic (eds.), *Global Catastrophic Risks* (Oxford: Oxford University Press).
- Joy, B. (2000) "Why the future doesn't need us", *Wired* 8.04.
- Kurzweil, R. (2005) *The singularity is near: when humans transcend biology* (New York: Viking).
- Lanier, J. (2000) "One-Half of a Manifesto", *Wired* 8 (21).
- Leslie, J. (1996) *The End of the World: The Science and Ethics of Human Extinction* (London: Routledge).
- Meadows, D. H., and Club of Rome. (1972) *The Limits to growth; a report for the Club of Rome's project on the predicament of mankind* (New York: Universe Books).
- Moravec, H. (1999) *Robot: Mere Machine to Transcendent Mind* (New York: Oxford University Press).

- Nordhaus, W. (2007) "A Review of the Stern Review on the Economics of Global Warming", *Journal of Economic Literature* forthcoming.
- Parfit, D. (1984) *Reasons and Persons* (Oxford: Clarendon Press).
- Pearce, D. *The Hedonistic Imperative* 2004. Available from https://www.hedweb.com/hedab.htm.
- Perrow, C. (1984) *Normal accidents: living with high-risk technologies* (New York: Basic Books).
- Posner, R. (2004) *Catastrophe: risk and response* (Oxford: Oxford University Press).
- Raup, D. M. (1991) *Extinction: bad genes or bad luck?* (New York: W.W. Norton).
- Rees, M. (2003) *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century - On Earth and Beyond* (Basic Books).
- Sagan, S. (2004) "The Problem of Redundancy Problem: Why More Nuclear Security Forces May Produce Less Nuclear Security", *Risk Analysis* 24 (4):935-946.
- Smil, V. (2006) *Transforming the twentieth century: technical innovations and their consequences* (Oxford: Oxford University Press).
- Solomon, S., Qin, D., Manning, M., et al. (2007) *Climate Change 2007: The Physical Science Basis. Contribution of the Working Group I to the Fourth Assessment Report.* Edited by Intergovernmental Panel on Climate Change (Cambridge: Cambridge University Press).
- Steinhardt, P., and Turok, N. (2002) "The Cyclic Universe: An informal introduction", *preprint* arXiv:astro-ph/0204479v1.
- Stern, N., and Great Britain Treasury (2006) *The economics of climate change: Stern review on the economics of climate change* (England: HM Treasury).
- Tainter, J. A. (1988) *The collapse of complex societies New studies in archaeology* (Cambridge: Cambridge University Press).
- Ulam, S. (1958) "John von Neumann 1903-1957", *Bulletin of the American Mathematical Society* (May).
- United Nations Population Division (2004) "World Population Prospects: The 2004 Revision", *Population Database.*
- Vinge, V. (1993) "The Coming Technological Singularity", *Whole Earth Review* Winter issue.
- Wolfers, J., and Zitzewitz, E. (2004) "Prediction markets", *Journal of Economic Perspectives* 18 (2):107-126.

- Wright, R. (1999) *Nonzero: The Logic of Human Destiny* (New York: Pantheon Books).
- Yudkowsky, E. (2007) "Artificial Intelligence as a Positive and Negative Factor in Global Risk", in Nick Bostrom and Milan Cirkovic (eds.), *Global Catastrophic Risks* (Oxford: Oxford University Press).