EXISTENTIAL RISKS FAQ

v. 1.2 (2013)
Nick Bostrom
University of Oxford
Future of Humanity Institute
www.existential-risk.org

Contents

What is an existential risk?

An existential risk is one that threatens the entire future of humanity. More specifically, existential risks are those that threaten the extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development. No existential catastrophe has ever occurred.

Human extinction would be an existential catastrophe if it happens before the heat death of the universe or before our potential for creating value has been fully realized. Some scenarios in which humanity survives would also be existential catastrophes if they involve a *permanent* and *drastic* destruction of humanity's future potential—something that is to humankind what a lifetime prison sentence or severe brain damage is to an individual.

"Humanity", in this context, does not mean "the biological species *Homo sapiens*". If we humans were to evolve into another species, or merge or replace ourselves with intelligent machines, this would not *necessarily* mean that an existential catastrophe had occurred—although it might if the quality of life enjoyed by those new life forms turns out to be far inferior to that enjoyed by humans.

What are the biggest existential risks?

Humanity's long track record of surviving natural hazards suggests that, measured on a timescale of a couple of centuries, the existential risk posed by such hazards is rather small. This finding is supported by direct analysis of specific hazards from nature.

The great bulk of existential risk in the foreseeable future is *anthropogenic*; that is, arising from human activity. In particular, most of the biggest existential risks seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology. As our powers expand, so will the scale of their potential consequences—intended and unintended, positive and negative.

For example, there appear to be significant existential risks in some of the advanced forms of synthetic biology, nanotechnology weaponry, and machine superintelligence that might be developed later this century. There might also be significant existential risk in certain future dystopian evolutionary scenarios, simulation-shutdown scenarios, space colonization races, nuclear arms races, climate change and other environmental disturbances, unwise use of human enhancement, and in technologies and practices that might make permanent global totalitarianism more likely.

Finally, many existential risks may fall within the category of "unknown unknowns": it is quite possible that some of the biggest existential risks have not yet been discovered.

How likely is it that humanity will succumb to an existential risk?

It is not possible to quantify rigorously the total level of existential risk. Estimates of 10-20% total existential risk in this century are fairly typical among those who have examined the issue, though such estimates rely heavily on subjective judgment. The real risk might be substantially higher or lower.

If technology carries existential risk, does that mean we should stop technological progress?

The answer is no, for several reasons. First, some technologies help reduce the existential risks created by other technologies or arising from nature. Second, the *permanent* failure to develop advanced technology would itself constitute an existential catastrophe, because the full realization of humanity's potential for creating and instantiating value requires advanced technology. Third, we might sometimes have reasons for action other than to minimize existential risk. Fourth, even a great effort by many people to halt technological progress would probably not succeed; and the disruption, conflict, or unilateral relinquishment that might result could easily increase the net level of existential risk. Fifth, there are more cost-effective means available to reduce existential risk.

There are particular technologies or applications that it makes good sense to try to stop or delay—biological weapons, for example. But in general, it is a difficult problem to figure out what kind of technology policy would be optimal from an existential-risk mitigation point of view.

Haven't people in the past often predicted the end of the world?

History is peppered with false prognostications of imminent doom. Blustering doomsayers are harmful: not only do they cause unnecessary fear and disturbance, but—worse—they deplete our responsiveness and make even sensible efforts to understand or reduce existential risk look silly by association.

To date, most doomsday prophets have not based their claims on science. It is therefore tempting to say that the solution is simply to distinguish superstition from science. However, although this distinction is important, it does not fully address the problem of doom-mongering. It is perfectly possible to produce overconfident science-based predictions of imminent catastrophe, or at least overconfident predictions that *appear* to be based on science. The predictions of Paul Ehrlich and the Club of Rome in the early 1970s might be viewed as examples of this. Furthermore, it is impossible to assess the likelihood of many of the biggest risks using strict and narrow scientific methods. There is no rigorously scientific way of foretelling how future technological capabilities will be used. Yet it would be an error to infer that powerful future technologies will pose no risk, or that we should focus our attention exclusively on those smaller risks that are easily quantifiable.

How does one study existential risks?

By and large, existential risks have barely been studied. We therefore know little about how big various risks are, what factors influence the level of risk, how different risks affect one another, how we could most cost-effectively reduce risk, or what are the best methodologies for researching existential risk.

Broadly, one can distinguish between studies that focus on one specific risk and ones that seek to illuminate a wide swath of existential risks. In the case of the former, the methodology will depend on which particular risk one is studying. Asteroid risk can be assessed on the basis of the distribution of impact craters from past events and by direct astronomical observation, supplemented with a damage model to estimate the consequences of an impact of a given magnitude. Climate change risk can be studied via climate simulations. Risks from future technologies might be studied by means of theoretical modelling to determine the capabilities enabled by various physically possible technologies, by examining what kinds of safeguards and countermeasures are feasible, and by considering the strategic context in which they will be deployed.

There are also some lines of investigation that promise to illuminate existential risk more generally. For example, one can study whether observation selection theory is applicable in some way to the assessment of net level of existential risk (such as via the Carter-Leslie Doomsday argument, considerations based on the Fermi paradox, or inferences from the simulation argument). One might also study human cognitive biases with the hope of finding ways of improving our intuitive judgments as they apply to existential risk. Other approaches to this issue also exist.

Why should I be concerned about existential risk?

A case can be made that our altruistic moral motivation should be focused on existential risk mitigation. To assess the value of reducing existential risk, we must assess the loss associated with an existential catastrophe. Hence we need to consider how much value would be realized

in the absence of such a catastrophe. It turns out that the ultimate potential for Earth-originating intelligent life is literally astronomical.

Even confining our consideration to the potential for biological human beings living on Earth gives a huge amount of potential value. If we suppose that our planet will remain habitable for at least another billion years, and we assume that at least one billion people could live on it sustainably, then the potential exists for at least 10^{16} human lives. These lives could be considerably better than the average contemporary human life, which is so often marred by disease, poverty, injustice, and various biological limitations that could be partly overcome through continuing technological and moral progress.

However, the relevant figure is not how many people could live on Earth but how many descendants we could have in total. One lower bound of the number of biological human life-years in the future accessible universe (based on current cosmological estimates) is 10^{34} years. Another estimate, which assumes that future minds will be mainly implemented in computational hardware instead of biological neuronal wetware, produces a lower bound of 10^{54} human-brain-emulation subjective life-years. (See "The Concept of Existential Risk" and "Astronomical Waste" for references and some further details.)

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

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

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

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

Maxipok is not a principle of absolute validity, since there clearly are moral ends other than the prevention of existential catastrophe. The principle's usefulness is as an aid to prioritization.

Shouldn't we focus on helping the people who exist now and need help, rather than on reducing existential risk?

The easy answer would be to say that we should do both. Perhaps the easy answer is the correct answer.

The underlying question hinges on deep and difficult issues in moral philosophy and population ethics—issues on which there is no consensus, even among smart and decent people who have thought long and hard about them. We should recognize that we are, for the time being, labouring under *moral uncertainty* on this point.

It is important to note, however, that *given* certain moral assumptions—assumptions that are widely, though by no means universally, accepted—existential risk mitigation by means of deontologically permissible methods is a dominant moral priority, as the answers to the previous questions illustrate.

Isn't this a very gloomy topic?

Perhaps, but many gloomy topics are pursued vigorously by many researchers, politicians, activists, and philanthropists—topics like war, human rights abuses, famine, educational deprivation, and disease. From one perspective, all of these areas are depressing. But from another perspective, they are also uplifting—particularly when we think of the great gains in human happiness that we have the ability to bring about by making progress on these problems. Likewise with existential risk: pondering catastrophic possibilities might be a downer, but thinking about how together we can help create a truly wonderful future for humankind and increase the chances of perhaps realizing unimaginably great values—this has the potential to be highly motivating, even uplifting.

If the field of existential risks mitigation has suffered from neglect and apathy, it is probably not because the topic is gloomy. Rather, part of the explanation might be because the topic can seem silly and/or impersonal. The topic can seem *silly* because the fact that there has never been an existential catastrophe makes the possibility of one seem far-fetched, because the biggest existential risks are all rather speculative and futuristic, because the topic has been besieged by doom-mongers and crackpots, and because there is as yet no significant tradition of serious scholars and prestigious institutions doing careful high-quality work in this area. The topic can seem *impersonal* because there are no specific identifiable victims—no heart-rending images of child casualties, for example. The main dangers seem to be abstract, hypothetical, and non-imminent, and to be the responsibility of nobody in particular.

What should be done to reduce existential risk?

There is probably much that could be done by societies and individuals to reduce net existential risk. Unfortunately, because the issue has scarcely been studied, our knowledge about what these potential risk-mitigation actions are—and which ones among them are most cost-effective—is very limited.

There are some obvious actions that would probably reduce existential risk by a tiny amount. For example, increasing funding for ongoing efforts to map large asteroids in order to check if any of them is on collision course with our planet (in which case countermeasures could

be devised) would probably reduce the asteroid risk by a modest fraction. Since—on a timescale of, say, a century—asteroids pose only a small existential risk, this is unlikely to be the most cost-effective way to reduce existential risk. Nevertheless, it might dominate conventional philanthropic causes in terms of expected amount of good achieved. (This is not obvious because conventional philanthropy likely has some *indirect* effects on the level of existential risk—for instance by changing the probability of future war and oppression, promoting international collaboration, or affecting the rate of technological advance.)

A somewhat more cost-effective project might involve operating a bunker or refuge that could enable a small human population to survive a wide range of catastrophic scenarios—plagues, nuclear winters, supervolcanic eruptions, asteroid impacts, complete collapses of human food production systems, and various "unknown unknowns". The refuge might be buried deep underground, stocked with supplies to last a decade or more, and designed to be easily defendable. Ideally it would be continually staffed by a quarantined population and stocked with tools that survivors could use in subsistence agriculture upon emerging from the shelter in the aftermath of a civilization-destroying catastrophe.

These two examples are given for illustration only. There are ideas for more targeted interventions that would probably be much more cost-effective, and additional ideas could be developed. This suggests an important point: *Research into existential risk and analysis of potential countermeasures is a strong candidate for being the currently most cost-effective way to reduce existential risk*. Such research involves, among other things, addressing certain methodological problems and strategic questions. Similarly, actions that contribute indirectly to producing more high-quality analysis on existential risk and a capacity later to act on the result of such analysis could also be extremely cost-effective. This includes, for example, donating money to existential risk research, supporting organizations and networks that engage in fundraising for existential risks work, and promoting wider awareness of the topic and its importance.

How can I help?

Everybody is in a position to help in some way. A small but useful contribution would be to help disseminate the key ideas, such as by linking to this website from webpages and blogs, translating the main papers into other languages, citing relevant work in academic articles and policy reports, covering the topic sensibly in the media, and so forth.

You can also contribute by funding individuals or organizations working on existential risk and related topics. Oxford University's Future of Humanity Institute is an academic research centre active in this area since 2006. FHI seeks to recruit the most brilliant minds and focus their attention on the most important problems. FHI also thinks about things like whether there are better things to do than to reduce existential risk, and about what methods one could use to answer this kind of question. Another organization that is seriously focused on existential risk reduction is the Machine Intelligence Research Institute. MIRI focuses on existential risks from machine superintelligence. There is an Existential Risk Reduction Career Network. These is also an effort currently underway to set up a Centre for the Study of Existential Risk at Cambridge University. Max Tegmark and others are founding the Future of Life Foundation, which is also intended to be actiev in this area.

For most people, the most effective way to contribute is probably by donating money, since that makes use of the principle of division of labour.