Founders Pledge

Safeguarding the Future Cause Area Report

AUTHOR:
JOHN HALSTEAD, DPHIL

LAST UPDATED 12/2020

Executive Summary

Homo sapiens have been on Earth for 200,000 years, but human civilisation could, if things go well, survive and thrive for millions of years. This means that whatever you value – be it happiness, knowledge, creativity, or something else –there is much more to come in the future. As long as we survive, humanity could flourish to a much greater extent than today: millions of generations could live lives involving much more happiness, knowledge, or creativity than today. Therefore, for members who value future generations, a top priority should be to safeguard the future of civilisation.

1. The Problem: emerging man-made risks

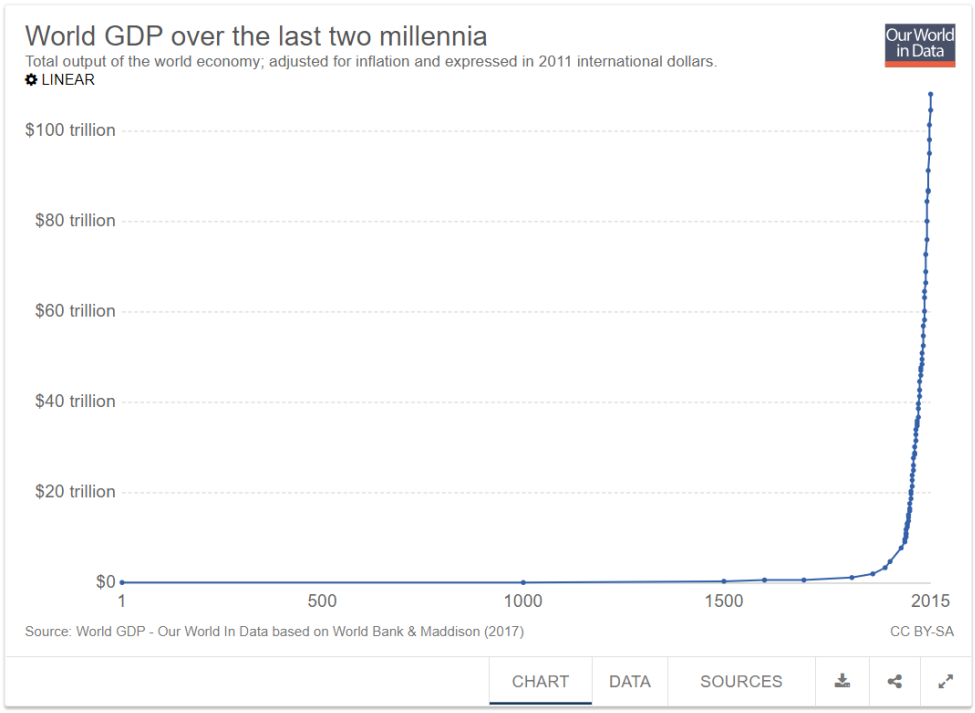
This is an especially urgent time to focus on safeguarding the future. *Homo sapiens* have survived for 200,000 years without being killed off by natural risks such as asteroids, and volcanoes, which is evidence that these pose a relatively small risk. However, the major risks we face today are *man-made*, stemming from our increasing power to affect our material conditions.

1.1. The Industrial Revolution and man-made risk

Following millennia of stagnation, innovation, automation and living standards exploded at the dawn of the Industrial Revolution.

Figure 1.

World GDP over the last two millennia



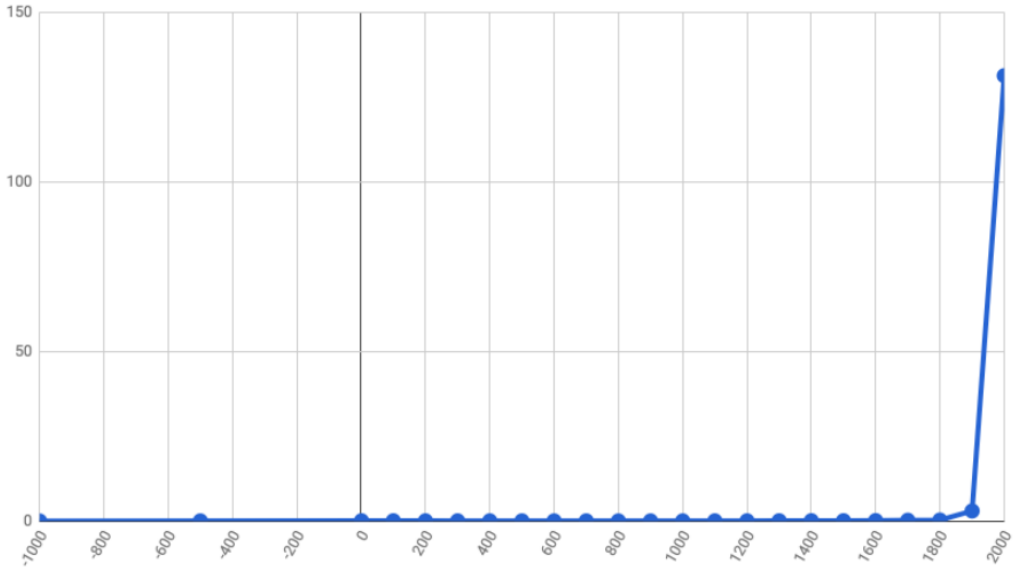
Source: Our World in Data, 'Economic Growth'

Since the Industrial Revolution, we gained the power to feed a growing population, to reduce child mortality, and to create technologies allowing us to travel and communicate across great distances.

However, our power to improve our material conditions increased in tow with our destructive power. According to work by Professor Ian Morris of Stanford, war-making capacity also exploded after the Industrial Revolution.

Figure 2.

Trends in war-making capacity in the last 3,000 years



Source: Luke Muehlhauser, 'How big a deal was the Industrial Revolution?' using data adapted from Morris, The Measure of Civilization, Princeton University Press (2013)[1]

The most dramatic shift in our destructive capacity came with the invention of nuclear weapons in 1945. This marked the dawn of a new epoch in which humanity for the first time potentially gained the ability to destroy itself. Developments in other areas may potentially be even more serious than nuclear weapons. Biotechnology and AI will greatly improve living standards, but according to many experts working in those fields, also carry potentially serious downside risks. Similarly, the burning of fossil fuels drove the huge increases in welfare we have seen over the last 200 years, but has caused CO_2 concentrations in the atmosphere to rise to levels unprecedented in hundreds of thousands of years, increasing the risk of extreme climate change.

Overall, the picture for the 21st century is one of increasing prosperity and flourishing, but also one of increasing risk that threatens to undo all this progress.

[1] Muehlhauser's post discusses the subtleties surrounding Morris' data. He cites Morris as saying "By "destructive power" I mean the number of fighters they can field, modified by the range and force of their weapons, the mass and speed with which they can deploy them, their defensive power, and their logistical capabilities."

1.2. Safeguarding the future is a highly neglected problem

Despite the unprecedented threat, global catastrophic risk reduction is highly neglected for several reasons. Future generations are the main beneficiaries of global catastrophic risk reduction, but they cannot vote, nor can they pay the current generation for protection. Global catastrophic risks are also global in scope, so no single nation enjoys all the benefits of reducing them.

Moreover, because the risks are unprecedented, increasing in the future, and also relatively unlikely, they are not salient to the public or to political leaders.[2] Consequently, leaders will tend to pay insufficient attention to them.

Finally, due to the psychological bias of scope insensitivity, people are insensitive to the large numbers at stake in global catastrophes. Our emotional reaction to finding out that a problem kills 1 million people or 100 million people is similar, and yet these tragedies call for very different social responses. The implications for global catastrophic risk are clear: there are trillions of potential lives in the future, but people may not take adequate account of this when thinking about the importance of global catastrophic risk.

For all these reasons, global efforts to safeguard the future have tended to be inadequate. For prospective donors, this means that the potential to find "low-hanging fruit" in this cause area is high at present. Just as VC investors can make outsized returns in large uncrowded markets, philanthropists can have outsized impact by working on large and uncrowded problems.

1.3. Overall risk this century

Estimating the overall level of global catastrophic risk this century is difficult, but the evidence, combined with expert surveys, suggests that the risk is plausibly greater than 1 in 100.[3] Given the stakes involved, we owe it to future generations to reduce the risk significantly.

[2] For a discussion of other biases relevant to the judgement of other existential risks, see Eliezer Yudkowsky, "Cognitive Biases Potentially Affecting Judgment of Global Risks," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (Oxford: Oxford University Press, 2008).
[3] For example, Toby Ord of the Future of Humanity Institute at Oxford puts the risk at around 1 in 12. Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury Publishing, 2020).

2. Outlining the major risks and potential ways forward

Based on expert surveys and our own reading of the evidence, we believe that the greatest threats to the flourishing of future civilisation stems from advances in biotechnology and advanced AI systems, with nuclear war and climate change also posing some risk.

2.1. Nuclear war

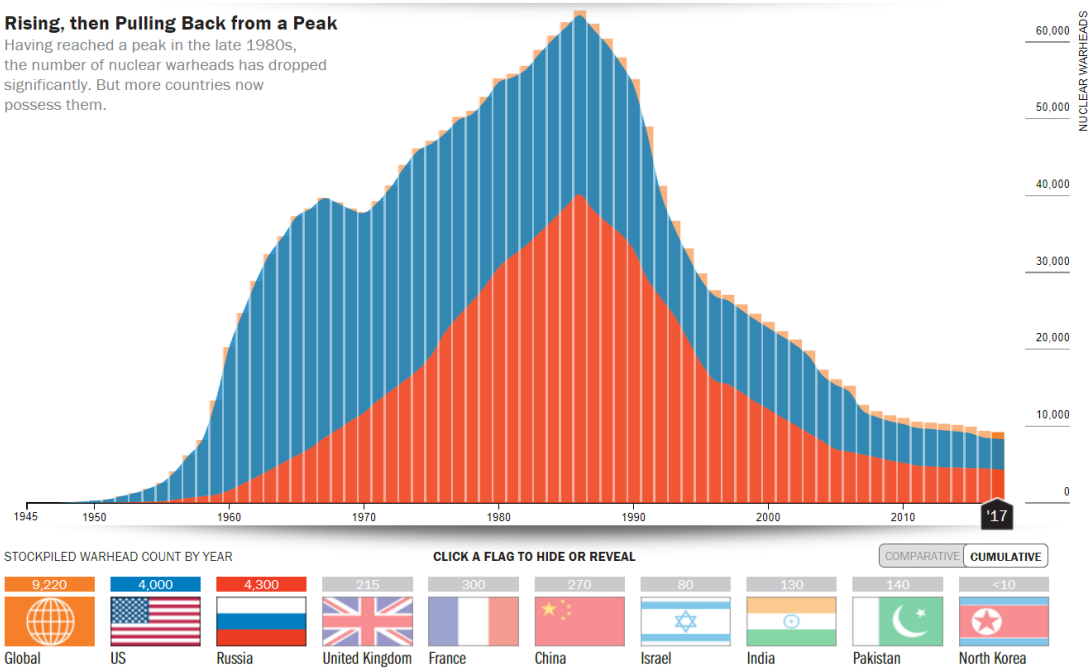
The discovery of nuclear weapons marked the dawn of a new epoch in which humankind may for the first time have gained the ability to destroy itself. The most concerning effect, first raised during the Cold War, is a potential *nuclear winter* in which the smoke from a nuclear war blocks out the Sun, disrupting agriculture for years. The potential severity of a nuclear winter is the subject of some controversy, but given the current split in expert opinion, it would be premature to rule it out.

As Figure 3 shows, global nuclear arsenals peaked in 1986 at around 64,000. While arsenals are significantly smaller today, each of the US and Russia together still have around 4,000 nuclear weapons each, with 1,400 of these strategically deployed (i.e. on ballistic missiles or at bomber bases).[4]

[4] Arms Control Association, "Nuclear Weapons: Who Has What at a Glance," June 2018, https://www.armscontrol.org/factsheets/Nuclearweaponswhohaswhat.

Figure 3.

Number of nuclear missiles held by the US and Russia



Source: Bulletin of the Atomic Scientists, Nuclear Notebook (2018

Reducing the risk of nuclear war

There are a number of possible ways to reduce the risk of civilisation-threatening nuclear winter.

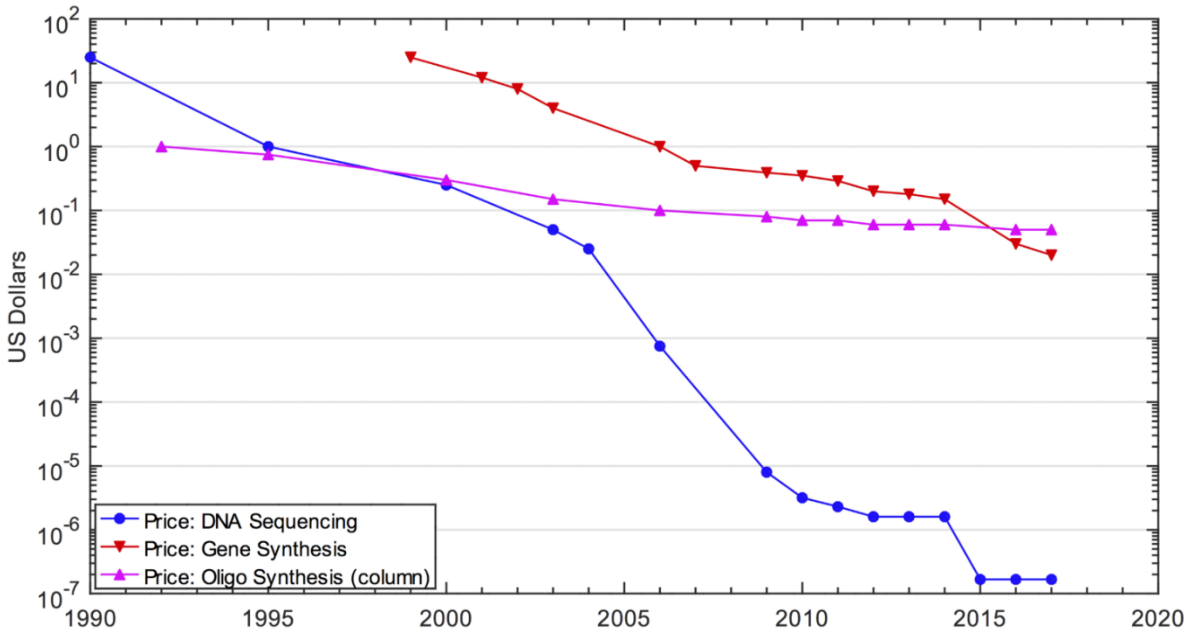
- Reduce the risk of conflict between major powers through diplomacy and other means.
- Change elements of nuclear strategy, such as taking nuclear weapons off hair-trigger alert.
- Reducing nuclear arsenals while maintaining the deterrence benefits of nuclear weapons. The US and Russian nuclear arsenals now far exceed what is needed to provide effective deterrence.
- Since much of the damage of nuclear war stems from smoke blocking out the sun, one could fund research into scaling up the production of food not reliant on sunlight.

2.2. Engineered bioweapons

Developments in biotechnology promise to bring huge benefits to human health, helping to cure genetic disease and create new medicines. But they also carry major risks. Scientists have already demonstrated the ability to create enhanced pathogens, such as a form of bird flu potentially transmissible between mammals, as well as to create dangerous pathogens from scratch, such as horsepox, a virus similar to smallpox. Figure 4 shows that the cost of gene synthesis has fallen by many orders of magnitude in recent years (note that the y-axis is a logarithmic scale).

Figure 4.

Cost of DNA sequencing, gene synthesis and oligo synthesis (oligos can be used to synthesise genes)



Source: Carlson, On DNA and transistors, (2016)

At present, the expertise and tacit knowledge required to exploit these improvements to create dangerous catastrophic biological events remain substantial. However, the worry is that as biotechnology capabilities increase and biotechnology becomes more widely accessible, scientists, governments or terrorists might be able, by accident or design, to create viruses or

bacteria that could kill hundreds of millions of people. Such weapons would be much harder to control than nuclear weapons because the barriers to acquiring them are likely to be considerably lower.

Reducing the risk of engineered pandemics

Various different approaches can be used to reduce the risk of engineered pathogens.

- Improve capacity for disease surveillance and response.
- Scenario planning for major global catastrophic biological risks, which would raise awareness about the risk and improve planning among important global actors.
- Investing in medical countermeasures, such as surge capacity for ventilators, vaccines, antivirals, and so on.
- Fostering a culture of safety among biotechnology researchers would also be valuable. Making researchers aware of the dual-use potential of research could allow researchers to produce beneficial insights without creating unnecessary risks.
- Developing and strengthening international biosafety norms to reduce the risk of accidental release from laboratories.

2.3. Artificial intelligence

Developments in artificial intelligence also promise significant benefits, such as helping to automate tasks, improving scientific research, and diagnosing disease. However, they also bring risks. Humanity's prosperity on the planet is due to our intelligence: we are only slightly more intelligent than chimpanzees, but, as Stuart Armstrong has noted, in this slight advantage lies the difference between planetary dominance and a permanent place on the endangered species list. Most surveyed AI researchers believe that we will develop advanced human-level AI systems at some point in the next 100 years. In creating advanced general AI systems, we would be forfeiting our place as the most intelligent being on the planet, but currently we do not know how to ensure that AI systems are aligned with human interests.

Experience with today's narrow AI systems has shown that it can be difficult to ensure that the systems do what we want rather than what we specify, that they are reliable across contexts, and that we have meaningful oversight. In narrow domains, such failures are usually trivial, but for a highly competent general AI, especially one that is connected to much of our infrastructure through the internet, the risk of unintended consequences is great. Developing a highly competent general AI could also make one state unassailably powerful, whichh increases the risk of misuse.

Managing the transition to AI systems that surpass humans at all tasks is likely to be one of humanity's most important challenges this century, because the outcome could be extremely good or extremely bad for our species.

Reducing the risk from advanced AI

There are several different ways to tackle the risks from advanced AI

- Build the field of AI researchers who are aware of and concerned about AI safety. This could be especially valuable to help build a culture of safety as AI systems develop over the coming decades.
- Technical research in computer science seems to have made progress in recent years,[5] and could be impactful if the timeline to advanced general AI turns out to be shorter than we think.
- Work on AI governance is in the early stages and could focus on researching the unique coordination challenges raised by transformative AI, and on advocating for awareness of these issues at the national and international level.

2.4. Climate change

Burning fossil fuels has allowed us to harness huge amounts of energy for industrial production, but also exacerbates the greenhouse effect. On current plans and policy, there is upwards of a 1 in 20 chance of global warming in excess of 6°C. This would make the Earth unrecognisable, causing flooding of major cities, making much of the tropics effectively uninhabitable, and exacerbating drought. Whether climate change is likely to cause a global catastrophe is unclear, and most of the risk seems to be very indirect. Donors interested in learning more about how to tackle climate change should see our Climate Change cause report and our Climate Fund.

[5] For an overview of recent developments, see footnote 15 in Robert Wiblin, "Positively Shaping the Development of Artificial Intelligence," 80,000 Hours, March 2017, https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence/. For discussion of some of the key issues in AI safety research, see the discussion by researchers at Google, OpenAI and Stanford in Dario Amodei et al., "Concrete Problems in AI Safety," *ArXiv:1606.06565 [Cs]*, June 21, 2016, http://arxiv.org/abs/1606.06565.

Acknowledgements

For helpful comments on this report, we are grateful to:

- Dr Seth Baum, Global Catastrophic Risk Institute
- Haydn Belfield, Cambridge Centre for the Study of Existential Risk
- Dr Niel Bowerman, 80,000 Hours
- Joe Carlsmith, Open Philanthropy Project
- Goodwin Gibbins, Imperial College London
- Howie Lempel, 80,000 Hours
- Dr Gregory Lewis, Future of Humanity Institute, Oxford
- Matthew van der Merwe, Future of Humanity Institute, Oxford
- Dr Stefan Schubert, University of Oxford
- Carl Shulman, Future of Humanity Institute, Oxford
- Dr Jess Whittlestone, Centre for the Future of Intelligence, Cambridge

Table of Contents

1. The Problem of Global Catastrophic Risk

Homo sapiens have been on planet Earth for 200,000 years, but human civilisation could, if things go well, survive and thrive for millions of years. This means that whatever you value — whether happiness, knowledge, or creativity — as long as we survive, there will be much more of it in the future. If we successfully navigate the various threats we face, our history so far will be a small fraction of the human story. Therefore, if our aim is to do as much good as possible, a top priority should be safeguard the future of civilisation. For pledgers who wish to benefit future generations, focusing on global catastrophic risk reduction looks a promising approach.

This is an especially urgent time to focus on safeguarding the future. Following millennia of stagnation, living standards improved enormously at the dawn of the Industrial Revolution, 200 years ago. We gained the power to feed a growing population, to reduce child mortality, and to create technologies allowing us to travel and communicate across great distances. However, our capacity to improve our material condition improved in tow with our capacity to create destruction. The most dramatic increase came with the invention of nuclear weapons at the end of the Second World War. This marked the dawn of a new epoch in which mankind may have gained the ability to destroy itself.

Nuclear war remains a risk today, but is now joined by other risks also driven by the vast improvement in our technology. Burning fossil fuels allowed us to harness huge amounts of energy, but also contributes to climate change. Developments in biotechnology and AI promise great benefits for human health and well-being, but also introduce novel and unprecedented risks.

In spite of the importance of the long term and the new context of emerging technological risk, reducing global catastrophic risk is highly neglected by governments and philanthropists

In the remainder of this section, we discuss the case for focusing on global catastrophic risk in more detail. An global catastrophic risk is defined here as a risk that threatens the premature extinction of sentient life or the destruction of its long-run potential.[6] On this definition, an global catastrophic risk need not kill everyone; anything that destroys humanity's long-term potential counts as a global catastrophic risk.

[6] For a similar definition, see Nick Bostrom, "Existential Risk Prevention as Global Priority," *Global Policy* 4, no. 1 (February 1, 2013): 15–31, https://doi.org/10.1111/1758-5899.12002.

1.1. How high is global catastrophic risk today?

Global catastrophic risk reduction is rarely held up as an important issue in political debate, nor is it on the agenda of many major philanthropists. In spite of this, the evidence suggests that the risk of a global catastrophe this century is greater than 1 in 100.

Human progress and anthropogenic risk

Humanity has survived for 200,000 years without being killed off, which is evidence that the baseline level of 'natural risk' from things such as asteroids and volcanoes is low.[7] However, natural disease risk is arguably exacerbated by certain features of modern society, such as increased international travel and population density. Thus, the historical record may not be a reliable guide to the threat posed by natural pandemics.[8] Nonetheless, the most serious global catastrophic risks this century are plausibly *anthropogenic* — driven by humanity — a product of our massively increasing technological capacity.

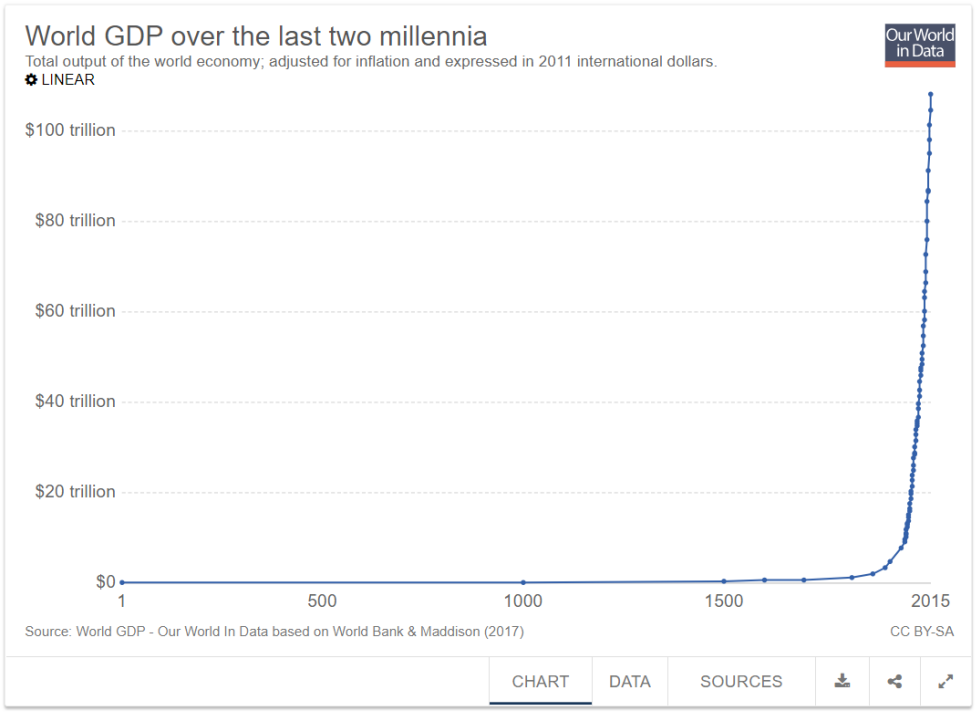
From the dawn of humanity until 1800, the rate of technological innovation across the globe had been stagnant, even as humanity moved out of hunter-gatherer societies into agricultural and pre-industrial societies.[9] Living standards were probably worse for the typical English person in 1700 than they were for some hunter-gatherer societies using Stone Age technology: many hunter-gatherers had a better diet and worked fewer hours.[10]

The long-term pattern of technological stagnation ended abruptly at the dawn of the Industrial Revolution in northern England in 1800 when innovation, automation and living standards exploded.

[7] Andrew E. Snyder-Beattie, Toby Ord, and Michael B. Bonsall, "An Upper Bound for the Background Rate of Human Extinction," *Scientific Reports* 9, no. 1 (2019): 11054.
[8] David Manheim, "Questioning Estimates of Natural Pandemic Risk," *Health Security* 16, no. 6 (2018): 381–390.
[9] See the discussion by Our World in Data of economic growth prior to the Industrial Revolution.
[10] Jared Diamond, "The Worst Mistake in the History of the Human Race," Discover Magazine, 1999, http://discovermagazine.com/1987/may/02-the-worst-mistake-in-the-history-of-the-human-race; Gregory Clark, *A Farewell to Alms: A Brief Economic History of the World* (Princeton, NJ: Princeton University Press, 2009), chap. 3.

Figure 1.1.

World GDP over the last two millennia



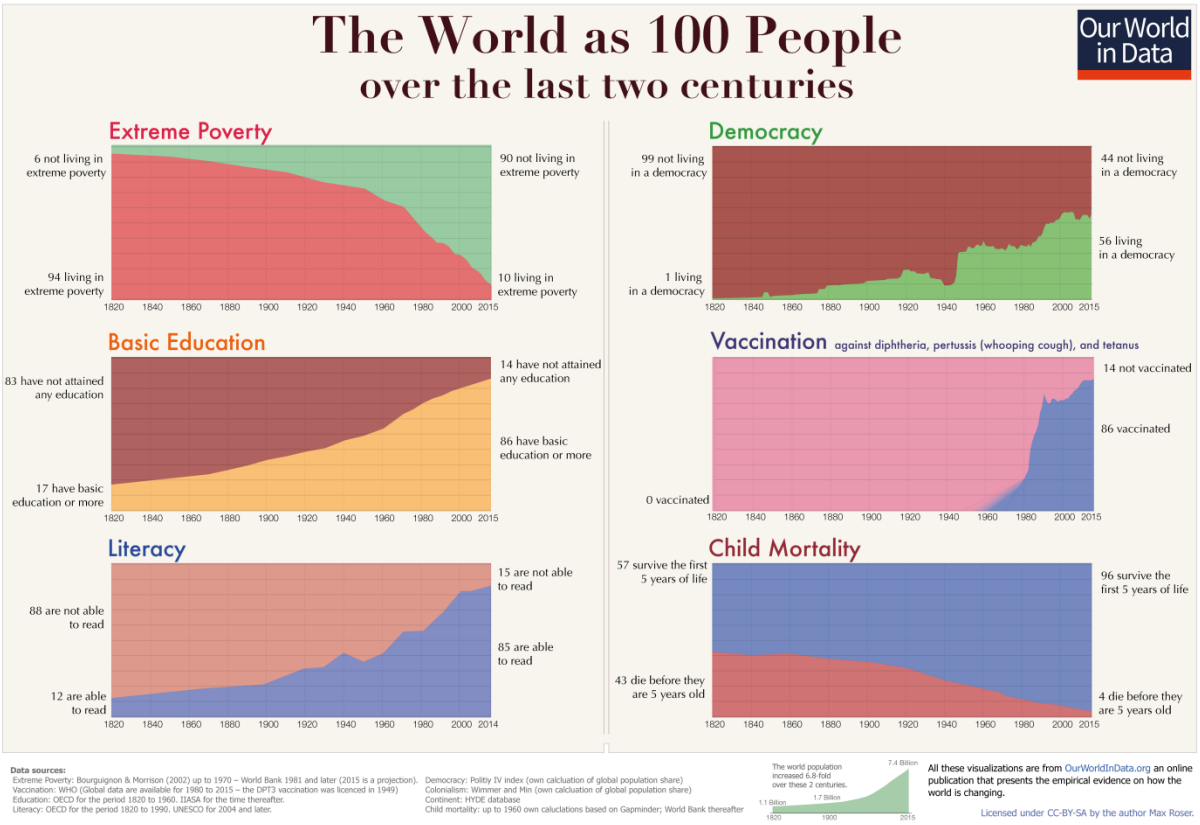
Source: Our World in Data, 'Economic Growth'

Practically all measures of human welfare suggest similarly explosive progress since 1800, following millennia of stagnation. Non-warfare violence has declined dramatically.[11] Illiteracy, child mortality and extreme poverty have all plummeted, while life expectancy and the percentage of people living in democracies has increased (see Figure 1.2).

[11] "Homicides," Our World in Data, accessed November 25, 2018, https://ourworldindata.org/homicides.

Figure 1.2.

Improvements in various measures of human welfare over the last 200 years

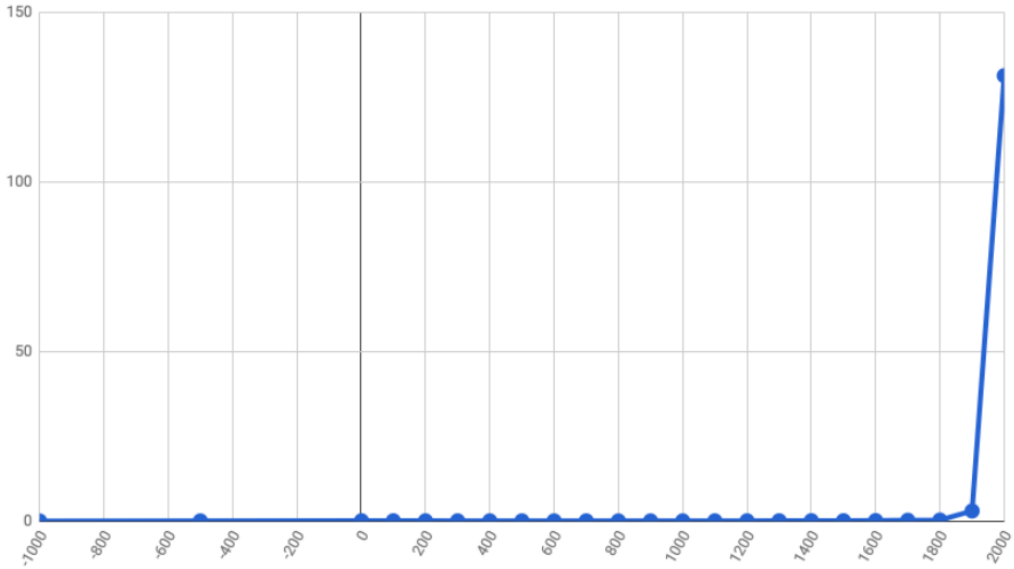


Source: Our World in Data, '[The short history of global living conditions and why it matters that we know it]'

However, our destructive power, or more specifically, our power to cause catastrophic damage, has increased hand-in-hand with our ability to improve our material situation. According to work by Professor Ian Morris of Stanford, war-making capacity has gone up in tow with living standards.

Figure 1.3.

Trends in war-making capacity in the last 3,000 years



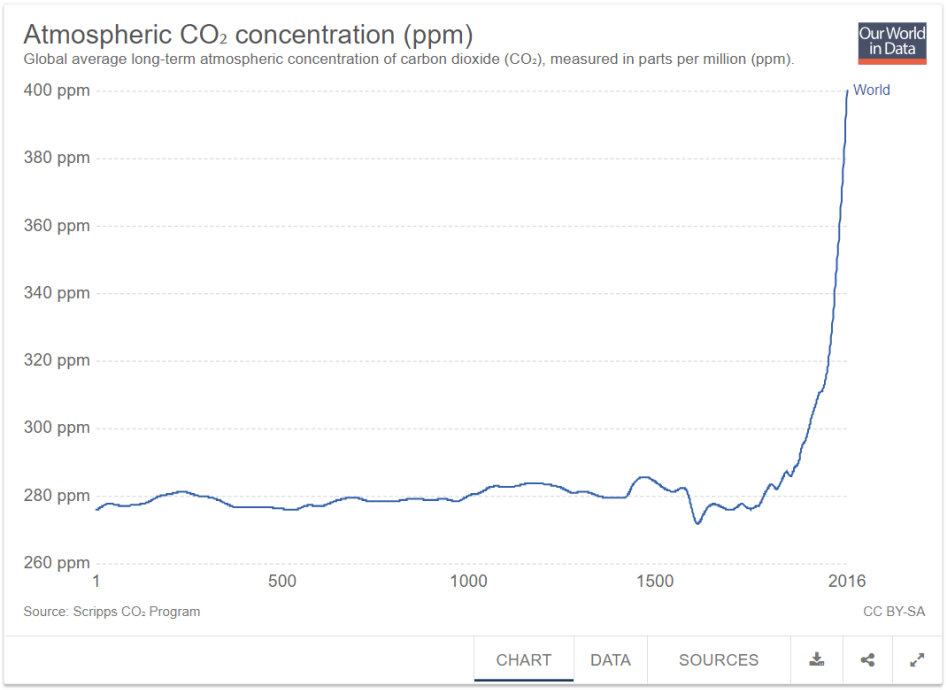
Source: Luke Muehlhauser, 'How big a deal was the Industrial Revolution?' using data adapted from Morris, The Measure of Civilization, Princeton University Press (2013)[12]

The most dramatic shift in our destructive capacity came with the invention of nuclear weapons in 1945. This marked the dawn of a new epoch in which humanity for the first time potentially gained the ability to destroy itself. Developments in other areas may potentially be even more serious than nuclear weapons. Biotechnology and AI will greatly improve living standards, but according to many experts working in those fields, also carry potentially serious downside risks. Similarly, the burning of fossil fuels drove the huge increases in welfare we have seen over the last 200 years, but has caused CO_2 concentrations in the atmosphere to rise to levels unprecedented in human history, increasing the risk of extreme climate change:

[12] Muehlhauser's post discusses the subtleties surrounding Morris' data. He cites Morris as saying "By "destructive power" I mean the number of fighters they can field, modified by the range and force of their weapons, the mass and speed with which they can deploy them, their defensive power, and their logistical capabilities."

Figure 1.4.

Concentrations of CO_2 in the atmosphere over the last 400,000 years



Source: NASA, Carbon Dioxide

The major anthropogenic risks are:

- Nuclear war
- Engineered pathogens
- Advanced AI
- Extreme climate change

We discuss each of these risks in more depth in section 2.

Estimating the combined risk from all of these threats involves highly subjective judgements. However, we do have access to expert judgements on nuclear war and advanced AI from which can deduce estimates of the total global catastrophic risk we face this century.

- **Nuclear war**

  The world appears to have come fairly close to nuclear war on a number of occasions in the past.[13] In a 2015 poll, 50 leading national security experts from across the world estimated the chance of a nuclear war between NATO and Russia of up to 4% in the next 20 years,[14] implying an 18% risk over the course of the next 100 years, if the risk remains constant. According to many scientists, billions could be threatened with starvation from an all-out nuclear war between the US and Russia with much of the Northern Hemisphere devastated. It is highly unclear whether society would ever fully recover. If, as seems fairly plausible, the probability that we fail to recover is at least 6%, then the global catastrophic risk from nuclear war alone is greater than 1% this century. Even if the chance we fail to recover is only 1%, the global catastrophic risk from nuclear war this century is still 1 in 555.

- **Advanced machine intelligence**

  In a 2017 survey of 352 machine learning researchers published at two of the leading machine learning conferences it was estimated that there is around a one in two chance that we will create an AI system that is better than humans at all relevant tasks by around 2060 and a 75% chance by the end of the century.[15] A subset of the researchers was asked about the outcomes of AI for humanity, and said that there is a 5% chance that the advanced machine intelligence would be "extremely bad" for humanity — an outcome equivalent to human extinction. Taking these estimates at face value implies that the chance of an global catastrophe caused by AI is 5%*75%, or 4%.

[13] Patricia Lewis et al., "Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy" (Chatham House, April 2014), https://www.chathamhouse.org//node/13981.
[14] PS21, "PS21 Survey: Experts See Increased Risk of Nuclear War," *PS21* (blog), November 12, 2015, 21, https://projects21.org/2015/11/12/ps21-survey-experts-see-increased-risk-of-nuclear-war/.
[15] Katja Grace et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *ArXiv Preprint ArXiv:1705.08807*, 2017.

Thus, deducing from expert estimates, the global catastrophic risk we face this century is plausibly upwards of 1 in 100.[16] This is roughly on a par with the lifetime risk of dying in a car accident for the typical European.[17] Moreover, this does not include other potentially important global catastrophic risks, such as engineered bioweapons, extreme climate change or currently unknown risks. Overall, the picture for the 21st century is one of increasing prosperity and flourishing, but also one of increasing risk that threatens to undo all this progress.

1.2. The ethics of safeguarding the future

The damage of a global catastrophe would be unprecedented and colossal. Even if we only focus on the current generation, billions of lives would be lost, exceeding deaths in the Second World War by two orders of magnitude. But another cost would be, as the philosopher Nick Bostrom puts it, that it would destroy the future.[18] The American astronomer Carl Sagan made this point eloquently when discussing the risk of nuclear winter:

> "If we are required to calibrate extinction in numerical terms, I would be sure to include the number of people in future generations who would not be born. A nuclear war imperils all of our descendants, for as long as there will be humans. Even if the population remains static, with an average lifetime of the order of 100 years, over a typical time period for the biological evolution of a successful species (roughly 10 million years), we are talking about some 500 trillion people yet to come. By this criterion, the stakes are one million times greater for extinction than for the more modest nuclear wars that kill "only" hundreds of millions of people.
>
> There are many other possible measures of the potential loss — including culture and science, the evolutionary history of the planet, and the significance of the lives of all of our

[16] Some experts who work on the area believe that the risk is much higher. For example, Toby Ord of the Future of Humanity Institute at the University of Oxford believes that the chance of an existential catastrophe this century is around 1 in 6. Robert Wiblin and Toby Ord, "Toby Ord - Why the Long-Term Future of Humanity Matters More than Anything Else," 80,000 Hours Podcast, accessed August 23, 2018, https://80000hours.org/podcast/episodes/why-the-long-run-future-matters-more-than-anything-else-and-what-we-should-do-about-it/.
[17] See the discussion by Bandolier.
[18] Bostrom, "Existential Risk Prevention as Global Priority," 17.

> ancestors who contributed to the future of their descendants. Extinction is the undoing of the human enterprise."[19]

The Oxford philosopher Derek Parfit made a similar case with the following thought experiment:

> "I believe that if we destroy mankind, as we now can, this outcome will be much worse than most people think. Compare three outcomes:
>
> 1) Peace.
> 2) A nuclear war that kills 99% of the world's existing population.
> 3) A nuclear war that kills 100%.
>
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is very much greater... The Earth will remain habitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second."[20]

Parfit is assuming here that the world would rebound from a nuclear war that kills 99% of the world's population, which is not obvious. But the general point that can be drawn from this is that, whatever you believe to be valuable, whether it be happiness, general human flourishing, freedom, justice, knowledge or art, there is much more to come in the future. As long as we survive, a far bigger and more advanced human civilisation could realise these values to a much greater extent than today: millions of generations could live lives involving much more happiness, flourishing, knowledge or art than today. This means that ensuring that we successfully navigate these emerging anthropogenic risks over the next 100 years would be extremely valuable, whatever your conception of the good life.

[19] Carl Sagan, "Nuclear War and Climatic Catastrophe: Some Policy Implications," *Foreign Affairs* 62, no. 2 (1983): 275, https://doi.org/10.2307/20041818.
[20] Derek Parfit, Reasons and Persons (Oxford: Oxford University Press, 1984), 453–54.

This also means that the expected value of global catastrophic risk reduction — which is the product of the probability of preventing a catastrophe and the impact of such an event — can be very high. The potential value to be gained is so large that even modest reductions in global catastrophic risk would be worthwhile. In standard risk analyses of ordinary projects, decisions about risk management are guided by the idea of expected value, so it seems that the same should apply for global catastrophic risk as well.

Although abstract, the idea that we should make some protections for the future is quite intuitive. Consider the following examples:

- Even though most of the costs of climate change will be felt by future generations, most people still think it worthwhile to reduce greenhouse gas emissions.
- When deciding how to safely store nuclear waste, governments have considered timeframes of hundreds of thousands of years.[21]
- For any form of pollution, the fact that it will damage humans hundreds of years in the future is not usually thought to be a reason to ignore that damage.
- Many people think we should make efforts not to use up non-renewable resources so that future generations can enjoy their benefits.

Justifications for protecting future generations often appeal to principles of *intergenerational equity* or *sustainable development*, according to which the needs and interests of future generations should get as much protection as those of the current generation.[22] Allowing global catastrophic risk to rise above 1% clearly shows insufficient concern for future generations, leaving them with a 1 in 100 chance of experiencing *none* of the benefits that we now enjoy. If we had a 1 in 100 chance of dying in a car accident, we would all make efforts to reduce the risk, for example by wearing a seatbelt. By the same token, it seems that we owe it to future generations to make extensive efforts to reduce global catastrophic risk down to an acceptable level.

[21] L. H. Hamilton et al., "Blue Ribbon Commission on America's Nuclear Future: Report to the Secretary of Energy" (Blue Ribbon Commission, 2012), 90.
[22] See for example, International Atomic Energy Agency, "Joint Convention on the Safety of Spent Fuel Management and on the Safety of Radioactive Waste Management," December 24, 1997, https://inis.iaea.org/search/search.aspx?orig_q=RN:36030798; Gru Brundtland et al., *Report of the World Commission on Environment and Development: Our Common Future* (Oxford University Press, 1987).

All this being said, ethical disagreement is pervasive, and we discuss possible ethical objections to the above arguments in section 1.5.

1.3. Global catastrophic risk is highly neglected

Global catastrophic risk reduction receives much less attention than is warranted. In spite of the level of risk today, the problem is rarely discussed by political leaders and is on the agenda of only a handful of philanthropists. As we discuss on our "How we think about charity" article, we think the neglectedness of a problem is a key determinant of how promising it is to work on. This is because, for uncrowded problems, the low-hanging fruit — the best opportunities for impact — are still available, and diminishing returns have yet to set in.

How neglected is global catastrophic risk?

All of the most pressing global catastrophic risks receive less attention than they should from governments and philanthropists, though some receive much more attention than others. The level of philanthropic neglectedness should be of particular interest to donors aiming to have an outsized impact.

- **Climate change**
  According to the UN Biennial Assessment of climate finance, around $900bn was spent on climate change in 2014.[23] However, much of this spending is less efficient than carbon pricing, the policy tool preferred by economists: carbon pricing schemes cover less than a quarter of total emissions and the carbon price is generally set below $10 per tonne of CO_2, far below the level recommended by economists.[24] For this reason, emissions have increased pretty much unchecked over the last few decades.[25] Climate change is the least neglected major risk among philanthropists. In 2018, major US philanthropists pledged $4bn

[23] UNFCCC Standing Committee on Finance, "Biennial Assessment and Overview of Climate Finance Flows," 2016, 56, http://unfccc.int/cooperation_and_support/financial_mechanism/standing_committee/items/10028.php.
[24] World Bank, "Carbon Pricing Watch 2017," 2017, 6–7.
[25] For an overview, see our climate change report.

up to 2023 to fight climate change,[26] a small fraction of the $410bn annually spent on philanthropy in the US,[27] but still much more than goes to the other major risks.

- **Nuclear security**
  US spending on nuclear weapons is in the tens of billions every year, with some unknown fraction of that devoted to safety, security and non-proliferation.[28] According to a report for the Hewlett Foundation, estimated spending on nuclear security by US philanthropists was around $31m in 2012.[29]

- **Engineered pathogens**
  In 2014, total US federal spending on biodefence was around $7bn,[30] though it is unclear how much of this was spent on the most extreme risks of engineered pathogens. Given that the US accounts for 24% of world GDP, we can roughly guess that between $5bn to $35bn is spent globally on reducing the risks of the most extreme pandemics.[31] It is unclear how much is spent by philanthropists on health security generally, but the area seems to have received renewed attention since the 2014 Ebola outbreak.[32] We would guess that the area now receives roughly as much attention as nuclear security (around $30m per year).

- **AI safety:**
  As of 2017, only $10m was spent on AI safety research,[33] though we think it is likely to receive much more in the future. Still, at least 1,000 times as much is spent on making AI

[26] Environment News Service, "Foundations Pledge $4 Billion Climate 'Down Payment,'" September 2018, http://ens-newswire.com/2018/09/26/foundations-pledge-4-billion-climate-down-payment/.
[27] Giving USA, "See the Numbers – Giving USA 2018 Infographic," accessed November 25, 2018, https://givingusa.org/see-the-numbers-giving-usa-2018-infographic/.
[28] "Budget | Department of Energy," accessed November 21, 2018, https://www.energy.gov/nnsa/budget.
[29] Redstone Strategy Group, "Clarifying Outcomes for the Nuclear Security Initiative," November 2012, 5, https://www.redstonestrategy.com/publications/clarifying-outcomes-nuclear-security-initiative/.
[30] Tara Kirk Sell and Matthew Watson, "Federal Agency Biodefense Funding, FY2013-FY2014," *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science* 11, no. 3 (September 2013): 196–216, https://doi.org/10.1089/bsp.2013.0047.
[31] World Bank, "GDP (Current US$)," accessed December 17, 2018, https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?year_high_desc=true.
[32] "Bill Gates Sees a Threat That Could Kill 30 Million People. Some Funders Are Paying Attention," Inside Philanthropy, accessed November 25, 2018, https://www.insidephilanthropy.com/home/2017/2/23/biosecurity-grants-open-philanthropy.
[33] Sebastian Farquhar, "Changes in Funding in the AI Safety Field," Centre for Effective Altruism blog, March 2017, https://forum.effectivealtruism.org/posts/Q83ayse5S8CksbT7K/changes-in-funding-in-the-ai-safety-field.

systems more competent.[34] Work on AI safety is divided between academic research institutes, and companies such as DeepMind.[35]

In total, with the exception of climate change, spending on the major global catastrophic risks by governments is probably less than $100bn, and spending by philanthropists probably now in the hundreds of millions. For context, global spending on high-end luxury goods in 2017 was $1.37tn.[36]

Why is global catastrophic risk so neglected?

The underinvestment in global catastrophic risk reduction is driven by a number of factors. Firstly, reducing global catastrophic risk is the responsibility of no single nation. The benefits of global catastrophic risk reduction are distributed across the globe, so a country that reduces global catastrophic risk enjoys only a fraction of the benefits but bears all the costs. Making humanity safe against global catastrophe is therefore a *global public good* that is subject to a classic *free rider problem:* countries have incentives to receive the benefits of risk reduction without contributing. This is why progress on climate change negotiations has been so slow. Each country has an incentive to let other countries reduce emissions while doing nothing themselves.[37]

Secondly, as mentioned above, most of the beneficiaries of global catastrophic risk reduction are future generations who neither have the political power to vote for their interests, nor the economic resources to compensate the current generation for protection. Thus, global catastrophic risk reduction is a *transgenerational public good*, meaning that it will tend to be underprovided by the current generation.

Several other factors plausibly also contribute to the neglect of global catastrophic risk prevention. Global catastrophes are by their nature unprecedented, which means that national and international governance regimes have not developed to manage them properly. In the same way, international governance institutions were not designed to deal with the novel threat of nuclear

[34] Steven Norton, "Artificial Intelligence Looms Larger in the Corporate World," Wall Street Journal, January 11, 2017, http://web.archive.org/web/20170126144740/http://blogs.wsj.com/cio/2017/01/11/artificial-intelligence-looms-larger-in-the-corporate-world/.
[35] Farquhar, "Changes in Funding in the AI Safety Field."
[36] Bain & Company, "Luxury Goods Worldwide Market Study, Fall–Winter 2017," Bain, December 22, 2017, https://www.bain.com/insights/luxury-goods-worldwide-market-study-fall-winter-2017/.
[37] William Nordhaus, "A New Solution: The Climate Club," The New York Review of Books, 2015, http://www.nybooks.com/articles/2015/06/04/new-solution-climate-club/.

weapons, and adequate consideration of key strategic concerns surrounding non-proliferation only came some time after the bombing of Hiroshima.

Moreover, because the risks are unprecedented, increasing in the future, and also relatively unlikely, they are not salient to the public or to political leaders.[38] Consequently, leaders will tend to pay insufficient attention to them.

Another important bias that could contribute to the relative lack of attention paid to global catastrophic risk is *scope neglect*.[39] People are naturally insensitive to differences in scale of many orders of magnitude. In a 1992 study, three groups were asked how much they would be willing to pay to save 2,000, 20,000 or 200,000 birds from drowning in uncovered oil ponds. The groups answered $80, $78 and $88, respectively.[40] Even though the numbers affected range over two orders of magnitude, this had little effect on the preferred response. In the same way, when we read the news, finding out that 10,000 people have died produces a similar emotional response as finding out that 100,000 people have died. The implications for global catastrophic risk are clear: there are trillions of potential lives in the future, but people may not take adequate account of this when thinking about the importance of global catastrophic risk.

1.4. Can we reduce global catastrophic risk?

It is intuitively difficult to see how to make progress on reducing global catastrophic risk and some believe this is a conclusive argument against working on it. However, there are historical examples of cases in which efforts to reduce global catastrophic risk have had some success. There also seem to be at least some clear ways to make progress on many of the risks in the future. Since the scale of the benefits would be so large, these reductions seem worthwhile. Indeed, while it is widely thought to be difficult to make progress on climate change, few people think that this is a reason for philanthropists to give up on the area. Given that other comparably important risks are much more neglected, the argument for working on them seems even stronger.

[38] For a discussion of other biases relevant to the judgement of other existential risks, see Yudkowsky, "Cognitive Biases Potentially Affecting Judgment of Global Risks."
[39] For discussion of scope neglect, see Daniel Kahneman et al., "Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues," *Journal of Risk and Uncertainty* 19, no. 1–3 (1999): 203–35.
[40] William H. Desvousges et al., "Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy," *Research Triangle Institute Monograph*, 1992.

Efforts to reduce the risks are of course high risk/high reward — most efforts will fail and only a handful will succeed. But this is true of all forms of political advocacy and of other areas, such as venture capital.

Historical successes

In the past, efforts to reduce some of the major risks we have discussed in this report seem to have had some success.

- **Nuclear weapons treaties:** The New START treaty, signed in 2010 by the US and Russia, limits the two countries to no more than 1,550 deployed nuclear warheads (warheads that are ready for active use) each. This is the latest of dozens of bilateral and multilateral arms pacts that have reduced US and Russian arsenals down from a combined peak of 60,000 to around 10,000 today.[41]

- **Climate treaties:** Progress on climate change prevention has been poor overall, but there have been some notable successes. Although flawed early on, as of November 2018, the EU's Emissions Trading Scheme placed a €20 price on each tonne of CO_2, making it one of the more stringent carbon pricing schemes.

- **The Global Seed Vault:** This frozen vault on an island between Norway and the North Pole contains the seeds of many important crop varieties, which would be useful in the event of an agricultural catastrophe.

- **Spaceguard:** Starting in the 1990s, a NASA-sponsored effort known as Spaceguard tried to track all Near Earth Objects (NEOs) — asteroids and comets — of more than 1 km in diameter.[42] If an object this size were to hit Earth, there could be a potentially civilisation-derailing catastrophe. Spaceguard has discovered more than 90% of these NEOs[43] showing the risk to be negligible in the next 100 years.[44] In this way, Spaceguard has greatly reduced the subjective risk associated with NEOs.

[41] Lawrence Korb, "Why It Could (but Shouldn't) Be the End of the Arms Control Era," *Bulletin of the Atomic Scientists* (blog), October 23, 2018, https://thebulletin.org/2018/10/why-it-could-but-shouldnt-be-the-end-of-the-arms-control-era/.
[42] Alan Harris, "What Spaceguard Did," *Nature* 453, no. 7199 (June 26, 2008): 1178–79, https://doi.org/10.1038/4531178a.
[43] See NASA, Discovery Statistics
[44] Harris, "What Spaceguard Did," fig. 2.

There are also cases in which individuals seem to have had a major effect on the risk of global catastrophes. As we discuss in section 2, during the height of the Cuban Missile Crisis, a Soviet naval officer, Vasili Arkhipov, plausibly played a crucial role in preventing nuclear war between the US and the USSR. There are a number of other (more arguable) cases in which individual judgement and discretion reduced the risk of nuclear war.[45] Scientists have also played a role in limiting the risk of global catastrophe. For example, as different countries raced to create the first nuclear bomb, individual scientists made extensive efforts to limit wider knowledge of the technology and the risk that the Nazis acquired it first.[46]

Future work

Looking to the future, there also seem to be promising ways to make progress on global catastrophic risk, which could focus narrowly on individual risks or more broadly on improving our general social and institutional ability to manage them.

- **Scientists working on emerging powerful technologies:** Like nuclear scientists, researchers working with emerging technologies such as autonomous weapons and engineered viruses will face momentous decisions in the future about whether to carry out and publish research with dual-use applications. If so, advocacy efforts to encourage a culture of safety and risk management could plausibly be high value. Research into technical AI safety, while embryonic, already appears to have produced results in limited domains, with the publication of a number of highly regarded papers on AI safety in the last few years.[47]
- **Climate change**: We provide two concrete recommendations of highly effective charities in our climate change report.
- **Our recommendations in this report**: We discuss concrete actions that philanthropists can take on the narrow risks in more detail in section 2 and provide recommendations on our Giving Recommendation page.

[45] For an overview, see Lewis et al., "Too Close for Comfort."
[46] Zia Mian, "Out of the Nuclear Shadow: Scientists and the Struggle against the Bomb," *Bulletin of the Atomic Scientists* 71, no. 1 (January 1, 2015): 59–69, https://doi.org/10.1177/0096340214563680.
[47] See footnote 15 of Wiblin, "Positively Shaping the Development of Artificial Intelligence."

Broad-based efforts to improve our ability to manage emerging threats also seem promising.[48] Work to improve global cooperation and institutional decision-making seem robustly good from the point of view of global catastrophic risk reduction; these efforts seem unlikely to backfire in any meaningful way. Improving our resilience to severe shocks to the food supply or to health systems also seems robustly positive.

1.5. Arguments against working to safeguard the future

The idea that global catastrophic risk is a high-impact problem to work on is counterintuitive. Few philanthropists focus on it and it is not a major topic of discussion in public debate. In part, we think this is due to the fact that there is very limited systematic focus on prioritising how to do the most good with your time and money. Among people who have thought in some depth about cause prioritisation, many have come to be convinced by the case for focusing on global catastrophic risk reduction.[49] Nonetheless, there are a number of possible objections to working on global catastrophic risk.

Ethical counterarguments

In section 1.2, we presented the ethical argument for the importance of protecting the long-term future of humanity, but there is considerable disagreement about ethics, even among moral philosophers whose sole job is to reason about ethics. We will now discuss some of these ethical objections.

Do future people matter?

Many philosophers and non-philosophers find appealing what is known as a *person-affecting view* of ethics, which says that adding possible future people to the world does not make the world better or worse.[50] This approach would appear to justify focusing on improving the welfare of the

[48] For more on this, see Benjamin Todd, "Why despite Global Progress, Humanity Is Probably Facing Its Most Dangerous Time Ever," 80,000 Hours, October 2017, 000, https://80000hours.org/articles/extinction-risk/.
[49] Outside of the effective altruism community, very little work is done on cause prioritisation from an impartial point of view. Surveys suggest that many leaders in effective altruism organisations, believe that existential risk reduction should be a top priority.
[50] See for example Jan Narveson and The Hegeler Institute, "Moral Problems of Population:," ed. Sherwood J. B. Sugden, *Monist* 57, no. 1 (1973): 62–86, https://doi.org/10.5840/monist197357134.

current generation. However, it is worth noting that given how many lives are at stake in the present generation only, there is still a lot at stake in global catastrophic risk reduction.[51]

Person-affecting views are held by many philosophers, but are subject to a number of counterarguments. Firstly, philosophers have found it difficult to state precisely exactly what is meant by person-affecting theories while remaining faithful to the intuition that motivated the theories.[52] This makes it difficult to yet have much confidence in any particular person-affecting view. Secondly, person-affecting theories have very counterintuitive implications in many cases. For instance, suppose that we could invest a small amount today to ensure that hazardous waste did not cause millions of deaths from cancer in the future. Person-affecting views are committed to saying that there is nothing good about making such an investment, yet most people would agree that we ought to make such an investment, even though doing so would only benefit possible future people.

Thirdly, the claim that possible future people do not matter morally implies that we should ignore future bad lives as well as future good ones.[53] Therefore, we should be indifferent about bringing into existence people with lives filled with intense suffering. Since proponents of person-affecting tend to find this unacceptable, they often defend a thesis called *the asymmetry*, which says that future bad lives should be taken into account, but not future good ones. However, this seems like an ad hoc adjustment — it is far from clear what motivates this asymmetry given the basic person-affecting intuition.[54] Moreover, asymmetric theories imply that if we have to choose between bringing into existence (a) someone with a life just barely worth living, and (b) someone with a very happy and flourishing life, we ought to be indifferent between (a) and (b). Again, this is difficult to accept.

As we have said, there is persisting disagreement about these ethical questions and settling them is very much a personal matter. If you are convinced of the person-affecting view, then it may be

[51] Gregory Lewis, "The Person-Affecting Value of Existential Risk Reduction," Effective Altruism Forum, April 2018, https://forum.effectivealtruism.org/posts/dfiKak8ZPa46N7Np6/the-person-affecting-value-of-existential-risk-reduction.
[52] On this, see Wiblin and Ord, "Toby Ord - Why the Long-Term Future of Humanity Matters More than Anything Else."
[53] For criticism of person-affecting views, see Hilary Greaves, "Population Axiology," *Philosophy Compass* 12, no. 11 (November 1, 2017): sec. 5, https://doi.org/10.1111/phc3.12442.
[54] Jeff McMahan, "Asymmetries in the Morality of Causing People to Exist," in *Harming Future Persons* (Springer, 2009), 49–68.

better to work on a problem affecting the current generation, such as malaria, or on global catastrophes that are likely to materialise in the next few decades.

Special obligations

One might also argue that even if reducing global catastrophic risk is extremely important, we have special obligations to our friends, family and to the current generation. However, these special obligations are compatible with the view that global catastrophic risk reduction should be a major priority. The arguments for focusing on global catastrophic risk are meant to show that *insofar as we are aiming to do the most good*, reducing global catastrophic risk looks a promising way forward, which is compatible with the view that there are other morally important things. Almost all moral viewpoints accept that doing good, impartially conceived, is a significant component of morality, and arguments about global catastrophic risk concern that component of morality.

Discounting

Economists often discount the future benefits deriving from a project by some factor, such as 3%. This approach makes sense for economic benefits because there are reasons to prefer having money sooner rather than later, such as that money can be invested and earn a return.

However, this is not a good reason to discount *intrinsically valuable* things that occur in the future, such as human happiness.[55] Economists often support discounting intrinsically valuable things on the basis that people in fact prefer to consume goods earlier in time. However, it is not clear why the fact that people discount *their own* welfare means that the welfare of future generations should be discounted. For example, just because I decide that I prefer a trip to the cinema today rather than next year, this doesn't mean that *someone else's* future welfare is worth less than their present welfare.

Discounting can have quite a large effect on judgements about the importance of future generations. Discounting welfare by 1% per year implies that the welfare of a person living 500 years in the future is worth less than 0.6% of that of a person today. We are not aware of any philosophers who believe that intrinsic goods such as human welfare should be discounted.

[55] For an overview of critiques of pure time discounting, see Hilary Greaves, "Discounting for Public Policy: A Survey," *Economics & Philosophy* 33, no. 3 (November 2017): 391–439, https://doi.org/10.1017/S0266267117000062.

The case for focusing on economic growth

Even if you accept the importance of future generations, there might be other ways to improve our long-term trajectory, aside from reducing global catastrophic risk. One possibility is focusing on increasing economic growth.[56] The effects of compounding growth can be substantial over the long term. For example, if the US' growth rate had been 1% lower between 1870 and 1990, then in 1990 the US would have been no richer than Mexico in 1990.[57]

However, reducing global catastrophic risk still seems like a higher priority than speeding up growth, for a few reasons. Firstly, broad economic growth and technological development are the source of progress, but also of the unacceptably high levels of risk that we face today. If protecting the future is important, pushing for fast development while incurring a greater than 1% chance of disaster is imprudent. It therefore makes more sense to take a more nuanced approach: rather than simply speeding up growth as much as possible, we should try to capture the benefits of growth while limiting the risks.[58] Secondly, advocacy for economic growth is much less neglected than work on global catastrophic risk. Economic growth is already one of the main aims of most domestic political parties and multilateral institutions, and there are already strong political pressures to push for higher growth. There is much less effort devoted to ensuring that growth is sustainable and safe.

Uncertainty about how to help

Given the complexity and diversity of the various risks, one could understandably remain uncertain about how best to help. The aim of this report is to resolve uncertainty for donors by providing concrete donation recommendations based on the judgement of experts with in-depth knowledge of the area and the recommended organisations. As protecting the long-term future is so important but so neglected, this is an especially good time for careful philanthropists to have an outsized impact.

[56] In *Stubborn Attachments*, Tyler Cowen defends the view that the main thing that matters is the *sustainable* rate of economic growth. Thus, he accepts the importance not creating unnecessary risk in the process of pursuing growth.
[57] Tyler Cowen, *Stubborn Attachments: A Vision for a Society of Free, Prosperous, and Responsible Individuals*, Kindle edition (Stripe Press, 2018), loc. 419.
[58] This is the idea of differential technological development, which is outlined here.

2. Outlining the major risks and potential ways forward

In this section, we provide an overview of the major, known global catastrophic risks that humanity faces this century: nuclear war, engineered bioweapons, advanced general artificial intelligence, climate change and various natural risks. We outline the level of the risk posed by different global catastrophic risks and discuss promising ways forward.

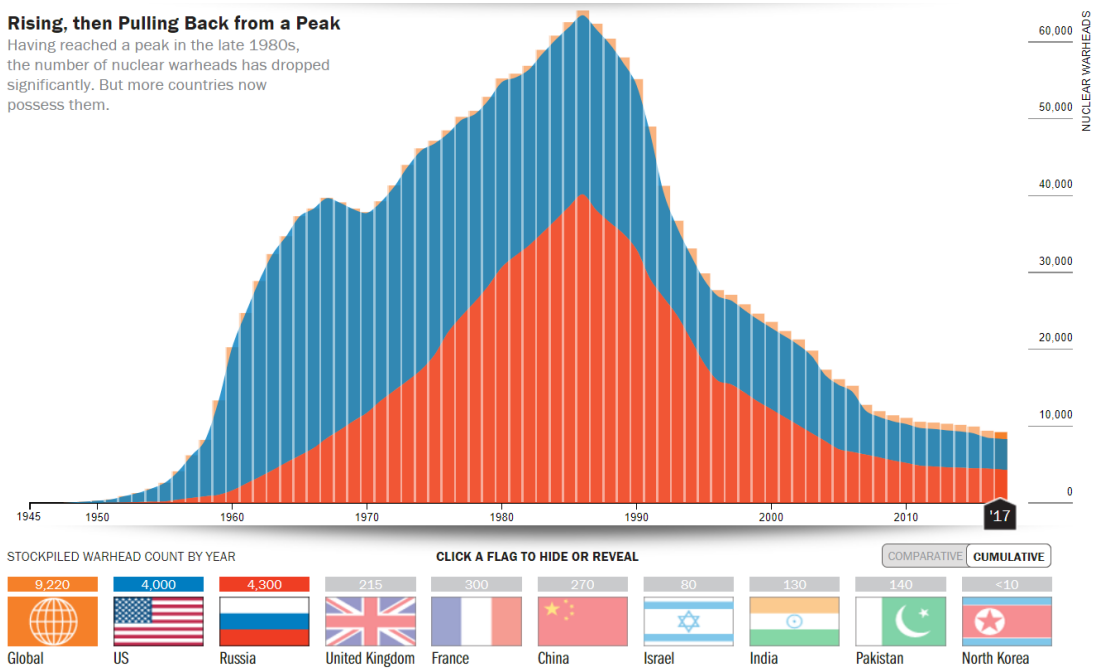
2.1. Nuclear war

The bombings of Hiroshima and Nagasaki demonstrated the unprecedented destructive power of nuclear weapons, with hundreds of thousands of people killed by only two nuclear bombs. Nuclear weapons have not been used since, but the risk of nuclear war was and remains non-negligible. As Figure 2.1 shows, global nuclear arsenals peaked in 1986 at around 64,000 warheads. While arsenals are significantly smaller today, the US and Russia together still possess 90% of global nuclear weapons, with around 4,000 each, although only 1,400 of these are strategically deployed (i.e. on ballistic missiles or at bomber bases) by each country.[59]

[59] Arms Control Association, "Nuclear Weapons: Who Has What at a Glance."

Figure 2.1.

Number of nuclear missiles held by the US and Russia



Source: Bulletin of the Atomic Scientists, Nuclear Notebook (2018

If these weapons were ever used in an all-out exchange, hundreds of millions of people would probably die in the blast, fire and radiation.[60] The catastrophic threat from such an exchange stems from the potential risk of *nuclear winter:* the burning of flammable materials sending massive amounts of smoke into the atmosphere, causing sustained global cooling even during summer, leading to massive agricultural disruption.

[60] Owen B. Toon, Alan Robock, and Richard P. Turco, "Environmental Consequences of Nuclear War," AIP Conference Proceedings 1596, no. 1 (May 9, 2014): 38, https://doi.org/10.1063/1.4876320.

Nuclear winter

The question of how severe a nuclear winter would be has been a matter of some debate among experts. The idea of nuclear winter came to prominence during the Cold War.[61] It proved controversial, with some scientists claiming that the risk had been overstated.[62] The idea received renewed interest with the publication of a number of studies using modern climate models to estimate the climatic effects of nuclear war. According to one study, an all-out exchange between the US and Russia involving around 4,000 weapons in total would put vast amounts of smoke into the atmosphere causing a drop in global temperatures of around 8°C for four to five years, making it impossible to grow food in most regions.[63]

If this were to happen, agricultural output would be severely undermined and some argue that billions would be threatened with starvation, with Toon *et al.* (2014) arguing that this would "likely eliminate the majority of the human population".[64] Whether a catastrophe of this kind would lead to permanent civilisational collapse is unclear. In the event of severe agricultural disruption, food stockpiles could potentially feed the global population for around four to seven months,[65] and people could take other extreme steps, such as slaughtering standing livestock and increasing fishing, which in all could feed the current population for perhaps one year.[66] Developing foods that are not reliant on sunlight also seems feasible,[67] and the incentives to take such steps would be strong.

Moreover, the Southern Hemisphere is largely free of nuclear weapons and so is unlikely to participate in a nuclear exchange, and certain countries, such as New Zealand and Australia would be spared the most extreme impacts and so would stand a better chance of survival, albeit in a

[61] For a brief overview, see Seth D. Baum, "Winter-Safe Deterrence: The Risk of Nuclear Winter and Its Challenge to Deterrence," Contemporary Security Policy 36, no. 1 (January 2, 2015): 125–27, https://doi.org/10.1080/13523260.2015.1012346.
[62] See for example John Maddox, "Nuclear Winter Not yet Established," Nature 308, no. 5954 (March 1984): 11, https://doi.org/10.1038/308011a0; Joyce E. Penner, "Uncertainties in the Smoke Source Term for 'Nuclear Winter' Studies," Nature 324, no. 6094 (November 1986): 222–26, https://doi.org/10.1038/324222a0.
[63] Alan Robock, "Nuclear Winter," *Wiley Interdisciplinary Reviews: Climate Change* 1, no. 3 (May 1, 2010): 421–22, https://doi.org/10.1002/wcc.45.
[64] Toon, Robock, and Turco, "Environmental Consequences of Nuclear War," 66.
[65] Carl Shulman, "What to Eat during Impact Winter?," *Reflective Disequilibrium* (blog), May 11, 2012, http://reflectivedisequilibrium.blogspot.com/2012/05/what-to-eat-during-impact-winter.html.
[66] David Charles Denkenberger and Joshua Pearce, *Feeding Everyone No Matter What: Managing Food Security after Global Catastrophe* (Amsterdam: Academic Press, 2015).
[67] Seth D. Baum et al., "Resilience to Global Food Supply Catastrophes," *Environment Systems and Decisions* 35, no. 2 (May 9, 2015): 301–13, https://doi.org/10.1007/s10669-015-9549-2.

greatly diminished state.[68] Some leading scientists working on nuclear winter have stated that the direct risk of complete extinction seems slim.[69] However, given the unprecedented nature of the catastrophe, it is unclear whether society would recover. In addition, society would be vulnerable to other catastrophes while in its weakened state.

Some research has suggested that even a smaller nuclear exchange involving only 100 nuclear weapons could produce a nuclear winter severe enough to threaten starvation for one billion people.[70] Catastrophes of this size do not, however, seem likely to threaten a global catastrophe. For example, 14th Century Europe bounced back quite easily from the Black Death, in which 30–50% of its population died,[71] even though it was much less technologically sophisticated than today. Similarly, other comparably large plagues have not threatened to cause the collapse of the societies affected. Thus, a US-Russia exchange seems the dominant concern.

Some of the science pertaining to a nuclear winter scenario has recently been criticised in a 2018 paper by Reisner *et al.*, which suggested that a small nuclear exchange would not cause a nuclear winter.[72] If the model in Reisner *et al.* is correct, the effects of an all-out exchange between the US and Russia have likely also been significantly overestimated.[73] Nevertheless, given the current split of opinion among experts in the literature, the global catastrophic risk of nuclear winter cannot be ruled out.

[68] Robock, "Nuclear Winter," 424.
[69] See for example Robock, 424; Carl Shulman, "Nuclear Winter and Human Extinction: Q&A with Luke Oman," Overcoming Bias, accessed November 2, 2018, http://www.overcomingbias.com/2012/11/nuclear-winter-and-human-extinction-qa-with-luke-oman.html; Malcolm W. Browne, "Nuclear Winter Theorists Pull Back," *The New York Times*, January 23, 1990, sec. Science, https://www.nytimes.com/1990/01/23/science/nuclear-winter-theorists-pull-back.html.
[70] Alan Robock and Owen Brian Toon, "Local Nuclear War, Global Suffering," *Scientific American* 302, no. 1 (2010): 74–81.
[71] Sharon N. DeWitte, "Mortality Risk and Survival in the Aftermath of the Medieval Black Death," *PLOS ONE* 9, no. 5 (May 7, 2014): e96513, https://doi.org/10.1371/journal.pone.0096513.
[72] Jon Reisner et al., "Climate Impact of a Regional Nuclear Weapons Exchange: An Improved Assessment Based On Detailed Source Calculations," *Journal of Geophysical Research: Atmospheres* 123, no. 5 (March 16, 2018): 2752–72, https://doi.org/10.1002/2017JD027331.
[73] The main points of contention concerns how much smoke would enter the atmosphere and how long it would remain in the upper atmosphere.

The probability of nuclear war

Many experts argue that the world has come close to nuclear war on a number of occasions,[74] though some dispute the severity of the alleged crises.[75] Perhaps the most serious incident occurred during the Cuban Missile Crisis in 1962 when US Navy ships dropped depth charges near a nuclear-armed Soviet submarine. Exhausted and thinking war might have already started, two of the officers on board argued that they should launch a nuclear warhead, which could have triggered all-out nuclear war. Authorisation of launch required the consent of the three most senior officers, of whom one, Vasili Arkhipov, refused to allow the launch.[76]

In 1983, at a time of heightened tensions, Soviet early warning systems showed five intercontinental ballistic missiles launching from the US. The duty officer at the time, Stanislav Petrov, was under orders to report any such launch to his superior officers, who would have eight to ten minutes to decide whether to respond.[77] Petrov decided not to do so, believing the launch to be a false alarm. It is possible that those higher in command would have come to the same judgement, but there is also a non-negligible chance that they would not have done.

Tensions have declined since the fall of the Berlin Wall, but the risk remains. Politics changes in highly non-linear and unpredictable ways, so it would be premature to rule out the possibility of nuclear war in the future. In a 2015 poll, 50 leading national security experts from across the world estimated the chance of a nuclear war between NATO and Russia of up to 4% in the next 20 years,[78] implying an 18% risk over the course of the next 100 years, if the risk remains constant. Other expert surveys also suggest that the risk is substantial.[79] Such polls are likely subject to significant subjective bias and selection effects, but at least suggest that the risk is non-negligible.

[74] See for example, Lewis et al., "Too Close for Comfort"; Union of Concerned Scientists, "Close Calls with Nuclear Weapons," 2015.
[75] Bruno Tertrais, "'On The Brink'—Really? Revisiting Nuclear Close Calls Since 1945," *The Washington Quarterly* 40, no. 2 (April 3, 2017): 51–66, https://doi.org/10.1080/0163660X.2017.1328922.
[76] Svetlana V. Savranskaya, "New Sources on the Role of Soviet Submarines in the Cuban Missile Crisis," *Journal of Strategic Studies* 28, no. 2 (April 1, 2005): 233–59, https://doi.org/10.1080/01402390500088312.
[77] Lewis et al., "Too Close for Comfort," 13.
[78] PS21, "PS21 Survey," 21.
[79] See for example Richard Lugar, *The Lugar Survey on Proliferation Threats and Responses* (United States Senator Richard G. Lugar, 2005).

Reducing the risk of nuclear war

There are a number of possible ways to reduce the risk of civilisation-threatening nuclear winter. Firstly, given the uncertainty and controversy surrounding nuclear winter, more research by diverse researchers on the science of nuclear winter would be valuable. Secondly, efforts could be made to reduce the risk of conflict between the major powers, such as the US, Russia and China, by encouraging dialogue and diplomatic outreach.

Thirdly, advocating for changing aspects of states' nuclear posture would reduce the risk of accidental war and miscalculation. For example, US nuclear weapons are currently on hair-trigger alert, with systems ready to launch within minutes of receiving a warning.[80] Many experts believe that weapons systems should be taken off hair-trigger alert, as this increases the risk of accidental war,[81] though other experts disagree.[82] Issues such as these are of course controversial, but illustrate that changing elements of US and Russian nuclear strategy could be a promising way forward.

The fourth approach seeks to limit the potential damage of a nuclear war by reducing nuclear arsenals while maintaining the deterrence benefits of nuclear weapons. The US and Russian nuclear arsenals now far exceed what is needed to provide effective deterrence, and nuclear arsenals should certainly be reduced at most to what is needed for effective deterrence. One recent report has suggested that the US only needs 650 weapons to have adequate deterrence.[83]

Whether arsenals need to be reduced further depends on how large an exchange would have to be to produce an global catastrophic risk-level nuclear winter. We argued above that an exchange of 100 weapons probably would not be sufficient to threaten human civilisation, but it is unclear what size arsenals would have to be to be 'nuclear winter-safe'. Various arms control treaties between the US and Russia have helped to greatly reduce each country's nuclear arsenals. The New START treaty, signed in 2010, limits the two countries to no more than 1,550 deployed nuclear warheads

[80] Bruce G. Blair, Jessica Sleight, and Emma Claire Foley, "The End of Nuclear Warfighting: Moving to a Deterrence-Only Posture" (Program on Science and Global Security, Princeton University; Global Zero, September 2018), chap. 6, https://www.princeton.edu/sgs/publications/articles/ANPR-Final-Blair.pdf.
[81] Blair, Sleight, and Foley, chap. 6.
[82] See for example, EastWest Institute, "Reframing Nuclear De-Alert: Decreasing the Operational Readiness of U.S. and Russian Arsenals," 2009, 7, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/reframing_dealert.pdf.
[83] Blair, Sleight, and Foley, "The End of Nuclear Warfighting: Moving to a Deterrence-Only Posture," 5.

(warheads that are ready for active use) each.[84] The treaty is set to expire in 2021, so advocacy for renewal or for a replacement treaty limiting arsenals further would be highly valuable.

Finally, we could focus our efforts on recovering from a nuclear winter. Since much of the damage of nuclear war stems from smoke blocking out the sun, the clearest way forward here would be to fund research into scaling up the production of food not reliant on sunlight.[85]

2.2. Natural and engineered pandemics

Pandemics involving engineered pathogens are thought by many experts to be one of the greatest global catastrophic risks in the 21st century. Here, we outline the emerging threat from natural and engineered pathogens.

The risk of natural pandemics

For most of human history, the greatest risk of mass fatalities has stemmed from natural pandemics. In the 1300s, the Black Death plague outbreak killed 30–50% of the European population.[86] The 1918 'Spanish flu' killed 50 million to 100 million people,[87] more people than died in World War One. These events are outliers, but history is punctuated by episodes of mass death from disease outbreaks.

However, there are some reasons to think that naturally occurring pathogens are unlikely to cause human extinction. Firstly, *Homo sapiens* have been around for 200,000 years and the *Homo* genus for around six million years without being exterminated by an infectious disease, which is evidence that the base rate of extinction-risk natural pathogens is low.[88] Indeed, past disease outbreaks have not come close to rendering humans extinct. Although bodies were piled high in the streets across

[84] Korb, "Why It Could (but Shouldn't) Be the End of the Arms Control Era."
[85] For discussion of this, see Denkenberger and Pearce, *Feeding Everyone No Matter What*. The non-profit ALLFED works on this issue.
[86] DeWitte, "Mortality Risk and Survival in the Aftermath of the Medieval Black Death."
[87] Niall Johnson and Juergen Mueller, "Updating the Accounts: Global Mortality of the 1918-1920 'Spanish' Influenza Pandemic," *Bulletin of the History of Medicine* 76, no. 1 (2002): 105–115.
[88] Toby Ord, "Will We Cause Our Own Extinction? Natural versus Anthropogenic Extinction Risks" (2014).

Europe during the Black Death,[89] human extinction was never a serious possibility, and some economists even argue that it was a boon for the European economy.[90]

Secondly, infectious disease has only contributed to the extinction of a small minority of animal species.[91] The only confirmed case of a mammalian species extinction being caused by an infectious disease is a type of rat native only to Christmas Island.[92] For humans, who are geographically dispersed, genetically diverse and capable of a rational response to problems, the risk therefore appears small.

Having said that, the context may be importantly different for modern day humans, so it is unclear whether the risk is increasing or decreasing. The risk may be increasing due to : (1) global travel, which allows more rapid pathogen spread; (2) increased population density; and (3) more close contact with animal populations due to both densely packed factory-farms and expansion into uninhabited areas lead to higher rates of emergence.[93] On the other hand, interconnectedness could also increase immunity by increasing exposure to lower virulence strains between subpopulations.[94] Moreover, advancements in medicine and sanitation limit the potential damage an outbreak might do. It is overall unclear how high the level of natural risk is today.[95] In our view the global catastrophic risk this century from natural pathogens is likely much lower than that of engineered pathogens, but the risk of natural pathogens may well have increased over the last 200 years. Our two recommended biosecurity charities each work to reduce the risk of both natural and engineered pathogens, so this debate does not appear to have important practical implications for the purposes of this report.

[89] Ole J. Benedictow, "The Black Death: The Greatest Catastrophe Ever," *History Today*, 2005, http://www.historytoday.com/ole-j-benedictow/black-death-greatest-catastrophe-ever.
[90] Gregory Clark, "Microbes and Markets: Was the Black Death an Economic Revolution?," *Journal of Demographic Economics* 82, no. 2 (2016): 139–165.
[91] Katherine F. Smith, Dov F. Sax, and Kevin D. Lafferty, "Evidence for the Role of Infectious Disease in Species Extinction and Endangerment," *Conservation Biology* 20, no. 5 (October 1, 2006): 1349–57, https://doi.org/10.1111/j.1523-1739.2006.00524.x.
[92] Kelly B. Wyatt et al., "Historical Mammal Extinction on Christmas Island (Indian Ocean) Correlates with Introduced Infectious Disease," *PLOS ONE* 3, no. 11 (November 5, 2008): e3602, https://doi.org/10.1371/journal.pone.0003602.
[93] Manheim, "Questioning Estimates of Natural Pandemic Risk," 385.
[94] Robin Thompson et al., "Increased Frequency of Travel May Act to Decrease the Chance of a Global Pandemic," *BioRxiv*, August 31, 2018, 404871, https://doi.org/10.1101/404871.
[95] For competing perspectives, see Snyder-Beattie, Ord, and Bonsall, "An Upper Bound for the Background Rate of Human Extinction"; Manheim, "Questioning Estimates of Natural Pandemic Risk."

Emerging risks from biotechnology

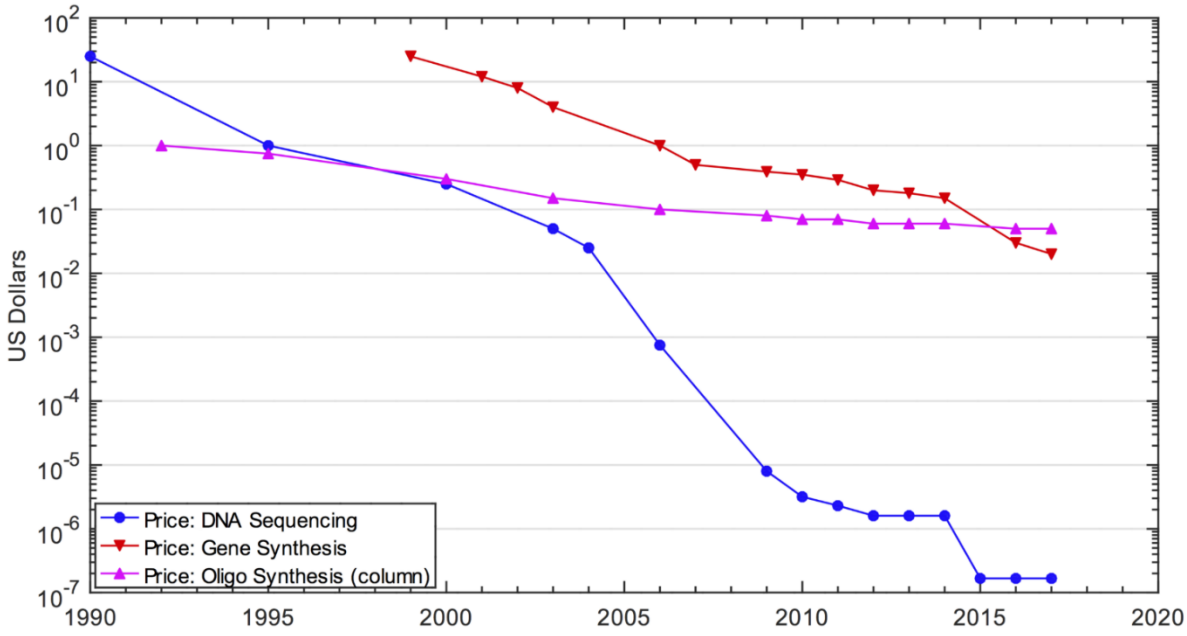
Improvements in biotechnology will bring great gains for human health, enabling us to cure genetic diseases, to create new vaccines, and make other important medical advances. However, biotechnology will also enable the features of pathogens to be determined by humans rather than by evolution, which could potentially greatly increase the probability of Global Catastrophic Biological Risks (GCBRs) — global catastrophes involving biological agents.[96] Researchers have, accidentally or otherwise, demonstrated the ability to design pathogens with dangerous new features. Controversial experiments published in 2012 detailed how to make a form of bird flu that is potentially transmissible between humans.[97] (Estimates of the case fatality rate in humans of bird flu range widely, from around 1% to 60%.[98]) In 2017, an American biotech company synthesised horsepox *de novo*, and similar techniques could potentially be applied to smallpox.[99] Biotechnology will enable us to better respond to pathogens such as these, but it seems plausible that in this area, attack is easier than defence.

If improvements in biotechnology continue on current trends, the cost and expertise required to produce dangerous pathogens will also continue to fall over the coming decades.[100] Figure 2.2 shows that the cost of gene synthesis has fallen by many orders of magnitude in recent years. However, the trend has slowed recently, and the expertise and tacit knowledge required to exploit these improvements remain substantial.

[96] Piers Millett and Andrew Snyder-Beattie, "Human Agency and Global Catastrophic Biorisks," *Health Security* 15, no. 4 (July 26, 2017): 335–36, https://doi.org/10.1089/hs.2017.0044.
[97] For discussion see Arturo Casadevall and Michael J. Imperiale, "Risks and Benefits of Gain-of-Function Experiments with Pathogens of Pandemic Potential, Such as Influenza Virus: A Call for a Science-Based Discussion," *MBio* 5, no. 4 (August 29, 2014): e01730-14, https://doi.org/10.1128/mBio.01730-14.
[98] Declan Butler, "Death-Rate Row Blurs Mutant Flu Debate," *Nature* 482, no. 7385 (February 13, 2012): 289–289, https://doi.org/10.1038/482289a.
[99] Gregory D. Koblentz, "The De Novo Synthesis of Horsepox Virus: Implications for Biosecurity and Recommendations for Preventing the Reemergence of Smallpox," *Health Security* 15, no. 6 (December 2017): 620–28, https://doi.org/10.1089/hs.2017.0061.
[100] Schoch-Spana et al., "Global Catastrophic Biological Risks."

Figure 2.2.

Cost of DNA sequencing, gene synthesis and oligo synthesis (oligos can be used to synthesise genes)



Source: Carlson, On DNA and transistors, (2016)

Creating a pathogen that would threaten human civilisation is impossible at present, but it is a real possibility that we will gain the ability at some point in the next century, as biotechnology improves.[101]

The possible sources of a future outbreak

Dangerous engineered pathogens could be released accidentally from lab research, or deliberately by terrorists or governments. Past data on the rate of accidental release of dangerous material from laboratories suggests that if there were widespread research on engineered potential

[101] Rees discusses the most threatening engineered viruses at Martin J. Rees, *Our Final Century: Will Civilisation Survive the Twenty-First Century?* (London: Arrow, 2004), 45–47.

pandemic pathogens, the future risk of accidental release would be decidedly non-negligible,[102] (though the size of the risk of accidental release is disputed).[103]

Secondly, the risk of deliberate bioterror attack seems to be the most concerning form of possible deliberate release. Aum Shinrikyo, a Japanese doomsday cult, successfully carried out a chemical weapons attack on the Tokyo subway in 1995.[104] Aum reportedly spent $10m on its bioweapons programme, which killed 12 people, but did not achieve its destructive aims, due to technical constraints and lack of expertise.[105] It is disconcerting to consider what a more qualified group could achieve with the technology available in the coming decades.

Finally, deliberate release for national military ends is potentially a concern. For instance,

> "In 1972, the United States, the Soviet Union and other nations signed the Biological and Toxin Weapons Convention that was supposed to ban biological weapons. At that very time, however, the Soviet Union was embarking on a massive expansion of its offensive biological weapons program, which began in the 1920s and continued under the Russian Federation at least into the 1990s."[106]

The Soviet biological weapons program was by far the largest and most sophisticated such programme ever undertaken by any nation. It was also intensely secretive, and was masked by layers of classification, deception and misdirection.[107] The risk of deliberate release of a GCBR-level pathogen by a state is arguably small enough to ignore given that states would have no incentive to ever deploy such a weapon, unless their own population was immune or had a cure. However,

[102] Marc Lipsitch and Thomas V. Inglesby, "Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens," *MBio* 5, no. 6 (December 31, 2014): 2, https://doi.org/10.1128/mBio.02366-14.
[103] For discussion see Ron A. M. Fouchier, "Studies on Influenza Virus Transmission between Ferrets: The Public Health Risks Revisited," *MBio* 6, no. 1 (February 27, 2015): e02560-14, https://doi.org/10.1128/mBio.02560-14; Marc Lipsitch and Thomas V. Inglesby, "Reply to 'Studies on Influenza Virus Transmission between Ferrets: The Public Health Risks Revisited,'" *MBio* 6, no. 1 (February 27, 2015): e00041-15, https://doi.org/10.1128/mBio.00041-15.
[104] For an excellent overview of Aum Shinrikyo's chemical and biological weapons programme, see Center for a New American Security, "Aum Shinrikyo: Insights Into How Terrorists Develop Biological and Chemical Weapons," December 2012, https://s3.amazonaws.com/files.cnas.org/documents/CNAS_AumShinrikyo_SecondEdition_English.pdf?mtime=20160906080510.
[105] Sonia Ben Ouagrham-Gormley, "Barriers to Bioweapons: Intangible Obstacles to Proliferation," *International Security* 36, no. 4 (April 1, 2012): 99, https://doi.org/10.1162/ISEC_a_00077.
[106] Steven Aftergood, "The History of the Soviet Biological Weapons Program," *Federation Of American Scientists* (blog), accessed October 12, 2018, https://fas.org/blogs/secrecy/2012/07/soviet_bw/.
[107] Aftergood.

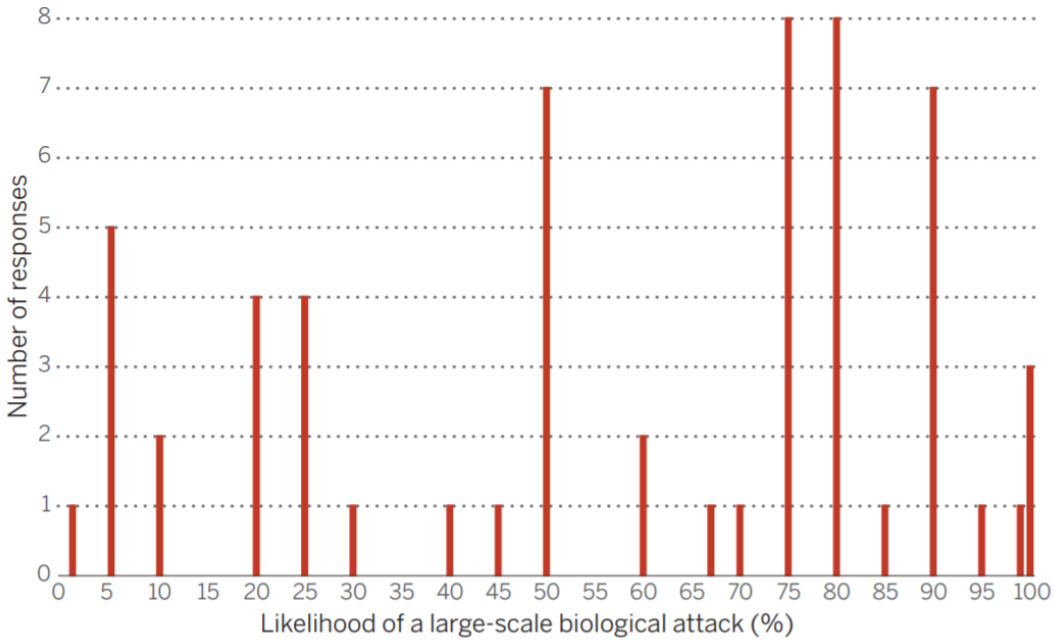
possession of a doomsday weapon could potentially be strategically useful, and the risk of miscalculation and accidental release would be a concern for any large bioweapons programme.

The likelihood of an engineered pathogen outbreak

It is very difficult to estimate the probability of the deliberate release of an engineered biological agent in the future. Figure 2.3 shows expert views on the probability of a large-scale biological attack involving engineered or natural pathogens between 2015 and 2025:

Figure 2.3.

Expert survey of the likelihood of a large-scale bioterror attack between 2015 and 2025



Source: Boddie et al., 'Assessing the bioweapons threat', Science (2015)

Figure 2.3 shows that there is significant disagreement among experts about the probability of an attack, with substantial fractions of experts giving extreme answers of upwards of 90% and less than 10%. For what it is worth, the mean estimate was 57.5%.[108] We do not put a great deal of

[108] Crystal Boddie et al., "Assessing the Bioweapons Threat," *Science* 349, no. 6250 (2015): 792.

weight on this and think that predictions such as these are often subject to bias and error,[109] but it is at the very least an indication that the risk of a deliberate biological attack cannot be ignored.

In line with the argument above, non-state actors were deemed to be the most likely source, with religious extremists and terrorists the most likely candidates, much more so than overt or covert use by a state.[110]

What can be done?

Various different approaches can be used to reduce the risk of engineered pathogens. Firstly, improving surveillance of and response to outbreaks would allow us to manage epidemic and pandemic outbreaks. The WHO's International Health Regulations are designed to ensure that national health systems have adequate surveillance and response of emerging health threats and that information on outbreaks is shared internationally. However, many poor countries cannot meet the regulations due to lack of capacity,[111] which suggests that additional funding and technical assistance would be beneficial.

Secondly, scenario planning for a GCBR pathogen would be highly valuable, but there have been relatively few to date.[112] Scenario exercises and plans allow key global actors to see the challenges that would be involved if there were a pandemic involving an engineered pathogen. Scenario planning would raise awareness about the emerging risk and would improve planning among important global actors.

The third broad way to reduce the risk of engineered pathogens is through medical countermeasures — either by ramping up spending on existing technology or by research and development into new technologies. Even rich countries do not currently have the capacity to deal with a major pandemic that infects tens of millions of people.[113] Therefore, investing in surge capacity for cheap ventilators and intensive care appears wise. There is also scope for research

[109] For an outstanding discussion of the problems of prediction, see Philip E. Tetlock and Dan Gardner, *Superforecasting: The Art and Science of Prediction*, Reprint edition (Broadway Books, 2016).
[110] Boddie et al., "Assessing the Bioweapons Threat," 792–93.
[111] Hans Kluge et al., "Strengthening Global Health Security by Embedding the International Health Regulations Requirements into National Health Systems," *BMJ Global Health* 3, no. Suppl 1 (January 1, 2018): e000656, https://doi.org/10.1136/bmjgh-2017-000656.
[112] One of the few examples that we are aware of is the CladeX exercise by the Johns Hopkins Center for Health Security, one of our recommended charities.
[113] John L. Hick et al., "Surge Capacity Principles: Care of the Critically Ill and Injured During Pandemics and Disasters: CHEST Consensus Statement," *Chest* 146, no. 4, Supplement (October 1, 2014): e1S-e16S, https://doi.org/10.1378/chest.14-0733.

and development into medical countermeasures, such as vaccines and broad-spectrum anti-virals, which would be useful in the case of a catastrophic biological event.[114]

Fourthly, fostering a culture of safety among biotechnology researchers would also be valuable. Making researchers aware of the dual-use potential of research could allow researchers to produce beneficial insights without creating unnecessary risks. A culture of safety could also deter research into techniques that could be exploited by malicious actors.

Finally, developing and strengthening international biosafety norms would help to reduce the risk of accidental release from laboratories. In the absence of agreed international norms, there is a risk that dangerous research could migrate to countries with lower biosafety standards.

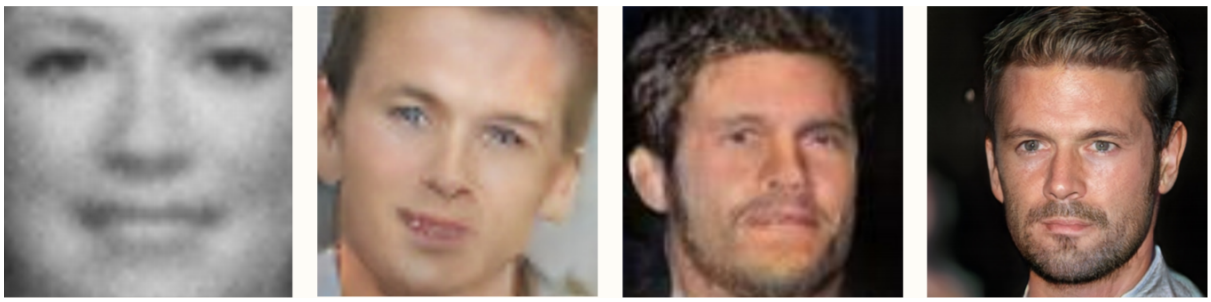
2.3. Advanced general artificial intelligence

In recent years, machine learning approaches to AI development have made strong progress in a number of domains. AI systems: now surpass humans at image recognition and games such as chess, Go and poker; have made huge progress in areas such as translation; and have even made novel scientific discoveries, such as predicting how proteins will fold. Figure 2.4 illustrates the rapid recent improvements in AI image generation: AIs are now able to produce synthetic images that are nearly indistinguishable from photographs, whereas only a few years ago the images they produced were crude and unrealistic.

[114] See the discussion by the Open Philanthropy project here.

Figure 2.4.

Increasingly realistic synthetic faces generated by variations on Generative Adversarial Networks



Source: Brundage et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation (2018): p.15. In order, the images are from papers by Goodfellow et al. (2014), Radford et al. (2015), Liu and Tuzel (2016), and Karras et al. (2017).

The number and breadth of the domains in which AI systems surpass human performance is sure to expand in the future and has raised the possibility that AI systems will one day outperform humans at all relevant tasks.

The transition to advanced general AI would be truly transformative for society, allowing us to solve many of our problems by improving medicine, transport, scientific research and so on. However, some prominent figures, such as Stephen Hawking and Elon Musk, along with many, though not all, AI researchers share these concerns. For instance, Stuart Russell, author of one of the leading AI textbooks is a leading advocate for concern about AI safety. In a large 2017 survey, leading AI researchers were asked to assign probabilities to outcomes following the development of human-level AI. The median probability was 20% for an "extremely good" outcome, and 5% for an "extremely bad" outcome (comparable to human extinction).[115]

A sketch of this worry is as follows. Humanity's dominance on the planet is entirely attributable to our intelligence, rather than to our strength or speed. As Stuart Armstrong, a Research Fellow at the Future of Humanity Institute, puts it:

> "The difference in intelligence between humans and chimpanzees is tiny, but in that difference lies the contrast between 7 billion inhabitants and a permanent place on the

[115] Grace et al., "When Will AI Exceed Human Performance?," 4.

> endangered species list. That tells us it's possible for a relatively small intelligence advantage to quickly compound and become decisive."[116]

In building advanced machine intelligence, we would forfeit our position as the most intelligent force on the planet, and we are currently doing so without a clear plan. Given the potential benefits we could enjoy if the transition to advanced general AI goes well, successfully navigating the transition to advanced AI seems to be one of the most important challenges we will face.

When will AI surpass humans at all tasks?

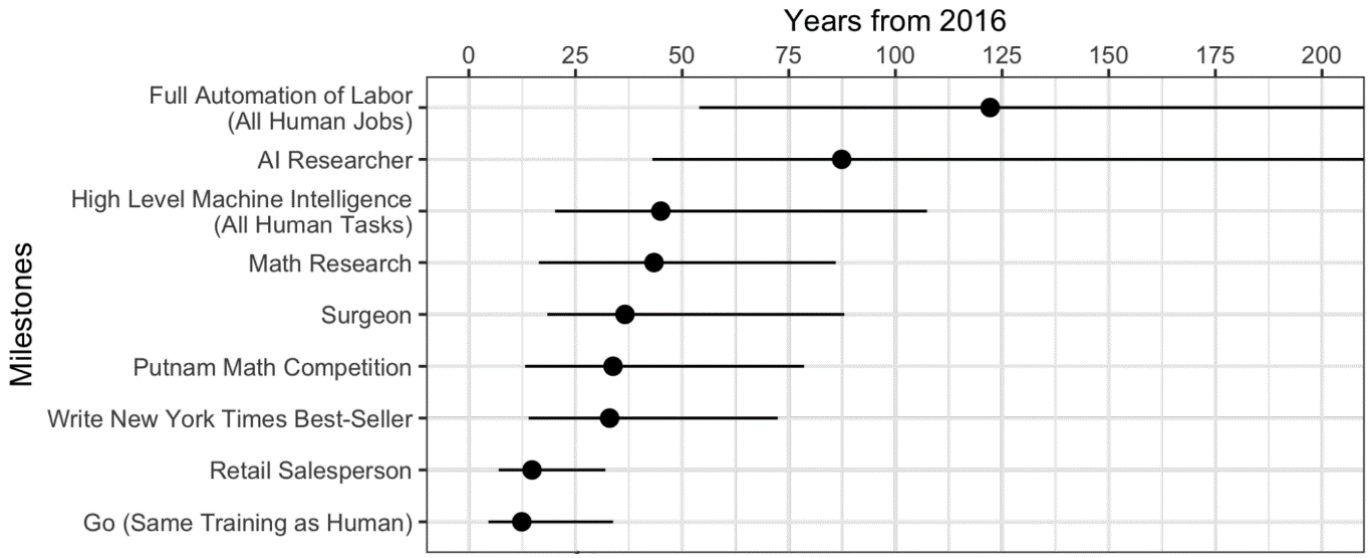
One natural thought about this is that AI will never actually reach and surpass the human level. However, in surveys, AI researchers tend to put a substantial probability on AI systems achieving human performance on most relevant tasks this century. In a 2014 survey of the 100 most-cited AI researchers (only 29 of whom responded), respondents gave a one in two chance of human-level AI systems by 2050, with AI systems probably surpassing humans 30 years after reaching the human level.[117]

In a larger survey by Grace *et al.* (2017), AI researchers gave very different answers depending on how an effectively identical question was framed: framing as a question about when all jobs would be fully automated produces a median estimate after 2100; but framing as a question about AI systems surpassing humans at all human tasks produces a median estimate of around 2060 (see Figure 2.5).[118]

[116] Ross Andersen, "Will Humans Be around in a Billion Years? Or a Trillion?," Aeon, February 2018, https://aeon.co/essays/will-humans-be-around-in-a-billion-years-or-a-trillion.
[117] Vincent C. Müller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion," in *Fundamental Issues of Artificial Intelligence*, ed. Vincent C. Müller (Berlin: Synthese Library; Springer, 2014).
[118] Grace et al., "When Will AI Exceed Human Performance?," 3.

Figure 2.5.

Timeline of median estimates for AI achieving human performance (with 50% confidence intervals)



Source: Grace et al., "When will AI exceed human performance? Evidence from AI experts" (2017)

These estimates are of course subject to the usual pitfalls involved in making subjective predictions about the future and may be subject to various forms of selection bias. Nonetheless, these surveys do suggest that the development of transformative AI in the next few decades cannot be ruled out. Even on the higher estimate produced by the "full automation" framing, researchers still estimated a 10% chance of full automation of all jobs by 2036.[119] Thus, the possibility of the invention of AI systems surpassing humans this century ought to be taken seriously.

AI accident risk

It is easy to see how nuclear war or engineered bioweapons could destroy humanity, but less clear how AI could do so. We briefly sketched the AI risk worry above, and in this section will discuss it in more detail. The basic concern is that, just as the fate of chimpanzees depends on the decisions of humans, if machine capabilities improve well beyond those of humans, and machines have

[119] Grace et al., 2.

increasingly general application and autonomy, then humans could come to depend entirely on the decisions and actions of machines. These decisions and actions could be hard to understand, predict and control, and might not be in our best interests. It is therefore important to ensure:[120]

- **AI value alignment**: The AI's goals are aligned with human interests.
- **Robustness**: The actions of the AI system are safe and reliable across different contexts.
- **Assurance**: We can understand and control AI systems during operation.

However, building safe, general and powerful AI systems that meet the three criteria above is likely to be difficult.[121]

Ensuring value alignment seems challenging because human values and interests are difficult to pin down: some of our most important concepts, such as love or happiness, are undefined and difficult to specify formally. This means that we may face a 'King Midas' problem — of getting exactly what we ask for, but not what we want. This problem is already prevalent in machine learning systems with narrow goals. For example:

- Researchers at OpenAI created an AI designed to complete a game called CoastRunners in which the aim is for a boat to finish a course as quickly as possible.[122] Because it was programmed to gain the most points, instead of doing this, the AI-controlled boat travelled in a circle hitting targets for points while repeatedly crashing and catching fire, rather than finishing the race. The video of its behaviour can be found here. [123]
- Researchers created an AI simulation designed to jump, but this was measured by how far it got its "feet" off the ground. Instead of jumping, the simulation learned to grow into tall vertical poles and do flips.

[120] On this, see DeepMind Safety Research, "Building Safe Artificial Intelligence: Specification, Robustness, and Assurance," *Medium* (blog), September 27, 2018, https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1.
[121] The obstacles to safe AI systems are discussed at length in Nick Bostrom, *Superintelligence : Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).
[122] "Faulty Reward Functions in the Wild," OpenAI Blog, December 22, 2016, https://blog.openai.com/faulty-reward-functions/.
[123] "Faulty Reward Functions in the Wild."

- OpenAI created a robotic arm trained to slide a block onto a target position on a table. The robotic arm sometimes does this by moving the table itself.

See this Google Doc for a list of other examples of AI systems doing what their creators specify, but not what they mean.

OpenAI discusses the wider significance of problems such as these:

> "It is often difficult or infeasible to capture exactly what we want an agent to do, and as a result we frequently end up using imperfect but easily measured proxies. Often this works well, but sometimes it leads to undesired or even dangerous actions. More broadly it contravenes the basic engineering principle that systems should be reliable and predictable."

These systems are all highly limited to narrow domains, but for highly competent and powerful general AIs, value misalignment could become a serious problem. Unless we ensure value alignment and robustness across contexts, for any goal we could give a highly competent AI, especially one that is connected to much of our infrastructure through the internet, the risk of unintended consequences seems great.[124] As Stuart Russell notes:

> "Any sufficiently capable intelligent system will prefer to ensure its own continued existence and to acquire physical and computational resources — not for their own sake, but to succeed in its assigned task."[125]

This means that intelligent machines could want to obscure its actions to prevent humans from interfering with its goals. For systems that are more intelligent than humans, ensuring adequate oversight could be difficult, because it could be increasingly difficult to understand what factors drove a particular decision,[126] and so might be hard to distinguish clever beneficial decisions from harmful ones. If something did start to go awry, the AI would have incentives to prevent us turning it off, and if the system was connected to the internet, turning it off could be difficult.

[124] See Bostrom, *Superintelligence*, chaps. 9–11.
[125] Stuart Russell, "Of Myths and Moonshine," Edge.org, 2015, https://www.edge.org/conversation/the-myth-of-ai#26015.
[126] Jacob Steinhardt, "Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems," *Academically Interesting* (blog), June 24, 2015, https://jsteinhardt.wordpress.com/2015/06/24/long-term-and-short-term-challenges-to-ensuring-the-safety-of-ai-systems/.

All of these technical issues in AI safety are the subject of ongoing research, and some have been solved in limited cases and domains.[127] However, research is very much in its infancy, and much more needs to be done to ensure the safety of highly competent AI systems.

Misuse risk

In addition to accident risk, there is the risk of deliberate misuse by states and other actors. The first state to develop and deploy a human-level AI system much before other countries would likely gain a decisive geopolitical and strategic advantage — extensive automation would massively increase growth and speed up innovation. This could in itself introduce global catastrophic risks because states could gain access to new weapons of mass destruction. Moreover, given the incentive to develop highly powerful general AI, if states fail to coordinate their efforts, there could be a "race to the bottom" in AI development and deployment as states skimp on safety in order to get better performance.[128]

Developments of AI for narrower applications could exacerbate risks in three areas:[129]

- **Digital security:** For example, AI could be used to automate cyberattacks, which are currently labour-intensive, and speech imitation technology could be used to manipulate people.
- **Physical security:** For example, AI systems could automate the tasks associated with drone attacks or subvert other physical systems, such as autonomous weapons and autonomous cars.
- **Political security:** AI could make surveillance, deception and propaganda easier for governments.

These risks are already present or imminent, so do not depend on the development of a human-level AI system. This suggests that there is a case for researchers to take account of the potential

[127] For an overview of major problems in AI safety, see Amodei et al., "Concrete Problems in AI Safety."
[128] Nick Bostrom, Allan Dafoe, and Carrick Flynn, "Public Policy and Superintelligent AI: A Vector Field Approach," in *Ethics of Artificial Intelligence*, ed. S.M. Liao (Oxford University Press, forthcoming), 7.
[129] For an overview, see Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *ArXiv:1802.07228 [Cs]*, February 20, 2018, http://arxiv.org/abs/1802.07228.

for misuse when creating research priorities, and for policymakers to work closely with researchers to minimise the risk of misuse.[130]

The neglect of safety and the case for forward planning

It might be argued that it is too soon to worry about transformative AI. However, even if advanced AI is only developed by the end of the 21st century, it would nevertheless be wise to put more effort into figuring out how to make safe AI systems, given the stakes. Indeed, the stakes are so great that transformative AI seems to be something we have to get right first time. Moreover, as we have seen, according to the (admittedly unstable and varied) judgements of experts in the field, there is a substantial probability of AI reaching or surpassing the human level in only a few decades.

At present, AI safety is highly neglected relative to its importance. While billions are spent on making AI more powerful, as of 2017, fewer than 100 people in the world were doing technical work on AI safety.[131] Global spending on AI safety — including both the governance and technical aspects — was only $9m in 2017.[132] Although we think this figure has likely increased substantially, we do not have a more recent estimate, and at least 1,000 times as much is spent on trying to improve AI capabilities.[133]

Consequently, the field of AI safety research is embryonic.[134] Nevertheless, progress does seem to have been made in the limited research so far. In a review of recent progress in technical AI safety, the impact research organisation 80,000 Hours writes:

> "The paper Concrete Problems in AI Safety, authored by machine learning researchers at Google, OpenAI, and Stanford, surveys a number of technical problems "that are ready for experimentation today and relevant to the cutting edge of AI systems" but are "likely to be robustly useful across a broad variety of potential risks, both short- and long-term."
>
> Another sign that there is low-hanging fruit in this space is the recent development of new formal frameworks and testing environments for studying safety problems in simple settings. These include safety environments at the OpenAI Gym, the logical

[130] Brundage et al., 4.
[131] Wiblin, "Positively Shaping the Development of Artificial Intelligence."
[132] Farquhar, "Changes in Funding in the AI Safety Field."
[133] In 2016, spending by companies alone (i.e. not included spending by governments) was $8bn, which is expected to grow to $47bn in 2020. Wiblin, "Positively Shaping the Development of Artificial Intelligence," fn 5.
[134] Steinhardt, "Long-Term and Short-Term Challenges to Ensuring the Safety of AI Systems."

> inductor and modal agent frameworks, cooperative inverse reinforcement learning, and algorithm learning by bootstrapped approval-maximization. Other recent progress on safety-relevant issues includes work on unsupervized risk estimation, reward engineering, and inverse reinforcement learning for bounded agents. The apparent rate of progress today suggests that more progress could be achieved if more research hours went into the problem."[135]

Although none of the technical problems have been solved for generally competent AIs, we now have a much clearer picture of the problem and approaches that might be worth exploring. Putting additional resources into the area now promises exceptional leverage.

Work on the governance aspects is also in the early stages, with a few academic research centres, such as the Future of Humanity Institute at the University of Oxford and the Cambridge Centre for the Study Existential Risk, devoted to governance questions. Work in this area could focus on researching the unique coordination challenges raised by transformative AI, and on advocating for awareness of these issues at the national and international level.

2.4. Climate change

Climate change is one of the most important problems facing humanity this century, but there are no comprehensive peer-reviewed studies of whether climate change is truly a global catastrophic risk, in the sense discussed here. The author of this report has, in an independent capacity, carried out a review of the evidence, which is available here.

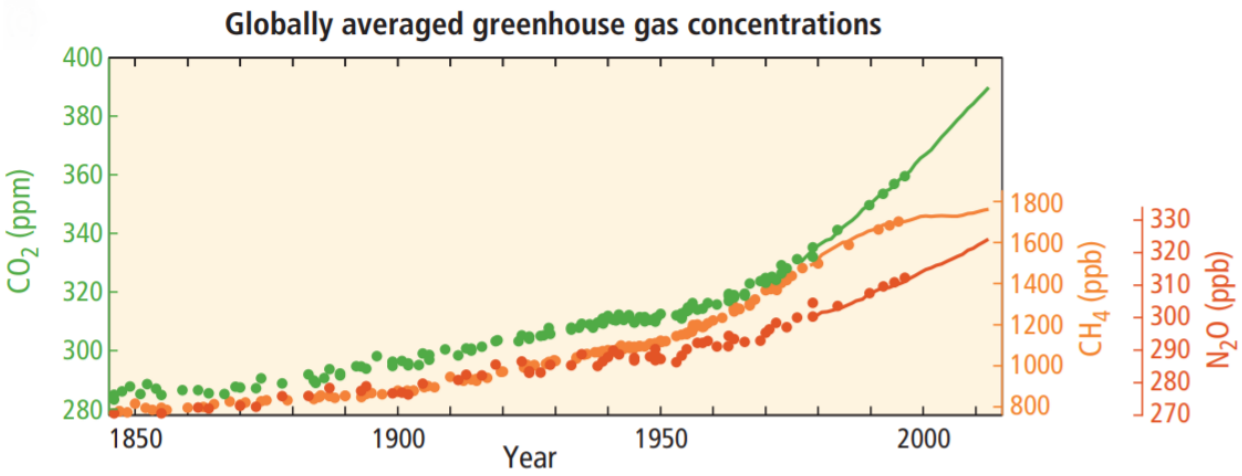
To understand the risk of climate change, it is important to first understand how the climate system works.[136] CO_2 emissions remain in the atmosphere for more than 1,000 years (unless we start to deliberately remove CO_2 from the atmosphere, which currently appears difficult and energy intensive).[137] This means that as long as CO_2 emissions are positive, CO_2 concentrations build up in the atmosphere. For most of human history, CO_2 concentrations in the atmosphere hovered around 280 parts per million (ppm). Due to the massive increase in the burning of fossil fuels and

[135] For an overview of recent developments, see footnote 15 in Wiblin, "Positively Shaping the Development of Artificial Intelligence." For discussion of some of the key issues in AI safety research, see the discussion by researchers at Google, OpenAI and Stanford in Amodei et al., "Concrete Problems in AI Safety."
[136] We provide a full overview of the area in our climate change report.
[137] Susan Solomon et al., "Irreversible Climate Change Due to Carbon Dioxide Emissions," *Proceedings of the National Academy of Sciences* 106, no. 6 (October 2, 2009): 1704–9, https://doi.org/10.1073/pnas.0812721106.

deforestation since the Industrial Revolution, in 2013 CO_2 concentrations passed 400 ppm for the first time in human history.[138]

Figure 2.6.

Atmospheric greenhouse gas concentrations 1850–2014



Source: Intergovernmental Panel on Climate Change (IPCC), Synthesis Report 2014, Figure SPM.1. (Parts per billion = ppb)

As this chart shows, other greenhouse gases, including methane and nitrous oxide, have risen along with CO_2. The warming effect of these gases is expressed in terms of CO_2-equivalent (CO_2e), which expresses the warming effect of greenhouse gases in terms of the functionally equivalent amount of CO_2. CO_2e concentrations today are around 445 ppm.[139]

In order to avoid extreme climate change, the policy challenge we face is to get to *net zero emissions*: at some point this century, there must be no emissions from power stations, factories, cars, trains, planes and ships, or we must start the expensive task of removing CO_2 from the

[138] NASA, "The Relentless Rise of Carbon Dioxide," Climate Change: Vital Signs of the Planet, accessed February 2, 2016, http://climate.nasa.gov/climate_resources/24/.
[139] "Atmospheric Greenhouse Gas Concentrations," Indicator Assessment, European Environment Agency, accessed November 29, 2018, https://www.eea.europa.eu/data-and-maps/indicators/atmospheric-greenhouse-gas-concentrations-10/assessment.

atmosphere. Reaching net zero emissions in the context of rapidly rising energy demand will be extremely challenging.

Tail risk

We are currently at around 445 ppm of CO_2e, but what will CO_2e concentrations eventually be? According to many sources, on current policy, we are headed to around 600–700 ppm of CO_2e by 2100.[140] However, this might be an underestimate of where we will ultimately end up because economic growth might be faster than expected, global political coordination might fail, and emissions might continue beyond 2100. Therefore, we cannot rule out eventually going above 1000 ppm.

How will the climate respond to emissions? The answer from climate science is that there is a worrying amount of uncertainty, with a substantial chance of extreme warming on plausible emissions scenarios. That is, climate change brings substantial *tail risk*.[141] Using UN Intergovernmental Panel on Climate Change estimates, the economists Wagner and Weitzman estimate the probability of at least 6°C (11°F) of warming, conditional on different levels of greenhouse gas concentrations:

[140] Joeri Rogelj et al., "Paris Agreement Climate Proposals Need a Boost to Keep Warming Well below 2 °C," *Nature* 534, no. 7609 (June 30, 2016): 635, https://doi.org/10.1038/nature18307.
[141] For a discussion of the importance of tail risk, see Gernot Wagner and Martin L. Weitzman, *Climate Shock: The Economic Consequences of a Hotter Planet* (Princeton: Princeton University Press, 2015).

Table 1.

The probability of >6°C of warming at different greenhouse gas concentrations[142]

CO_2e concentration (ppm)	400	450	500	550	600	650	700	750	800
Chance of >6°C (11°F)	0.04%	0.3%	1.2%	3%	5%	8%	11%	14%	17%

According to this research, even if we only end up at 500 ppm of CO_2e (which seems highly optimistic)[143] the probability of more than 6°C of warming is 1.2%. If we end up at 700 ppm, the chance of that would be 11%. Below, we discuss the negative consequences of 6°C of warming, which make a compelling case for strong action on climate change. Furthermore, even more extreme outcomes are possible. If we go past 1120 ppm, there is a greater than 66% chance of warming of between 3°C and 9°C, and at least a 2% chance of 12°C of warming.[144]

Much discussion of climate change conceals the tail risks of warming. For example, at the Paris Agreement, the world agreed to keep global warming below 2°C. However, in practice the world agreed to emit enough to stay within a "2°C carbon budget", which means "the amount of carbon we can emit to have a >66% chance of staying below 2°C".[145] So, even if we stay within our "2°C carbon budget", there would still be up to a one in three chance of going past 2°C.

[142] Wagner and Weitzman, 54.
[143] CO_2 concentrations are currently increasing by about 2ppm every year, so we will end up at 500ppm in around 25 years, on current trends.
[144] This follows from estimates of equilibrium climate sensitivity, which suggests that any doubling of CO_2 concentrations has a >66% chance of producing between 1.5°C and 4.5°C, and a 1-10% chance of more than 6°C.
[145] Joeri Rogelj et al., "Differences between Carbon Budget Estimates Unravelled," *Nature Climate Change* 6, no. 3 (March 2016): 245–52, https://doi.org/10.1038/nclimate2868.

The direct impact of extreme warming

There are two ways in which extreme warming might be thought to cause a global catastrophe: either directly via its environmental effects, or indirectly through causing global political instability and conflict. We discuss the direct effects in this subsection and indirect effects in the next.

The impacts of extreme warming are understudied. In spite of the fact that on current policy there is an 11% chance of more than 6°C of warming, very few studies reviewed in the 2014 IPCC *Impacts* report investigate the impact of warming of more than 4°C on crops, ecosystems, health, poverty, security or the economy.[146]

We now briefly review the evidence on the impact of extreme warming. An important factor to consider throughout this section is the timescale of extreme warming. Extreme warming will take several centuries to occur, as the additional heat from the greenhouse effect is absorbed by the ocean.[147] If so, we have lots of time to adapt to extreme warming.[148]

Sea level rise

On the highest emissions scenario considered by the IPCC, sea level is projected to rise by around 1 metre by 2100,[149] and by upwards of 10 metres over the course of millennia.[150] While this would be extremely bad, destroying most currently existing coastal cities, we would have lots of time to adapt by building flood defences and moving inland. Sea level rise would serve to shrink the habitable space on Earth, but would not by itself come close to threatening extinction.

[146] David King et al., "Climate Change–a Risk Assessment" (Centre for Science Policy, University of Cambridge, 2015), 46, www.csap.cam.ac.uk/projects/climate-change-risk-assessment/.
[147] IPCC, *Climate Change 2013: The Physical Science Basis: Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ed. T. F. Stocker et al. (Cambridge University Press, 2013), 1102–3.
[148] One possible exception to this is that the release of vast amounts of methane from clathrates, which could lead to rapid warming. This possibility has been posited by Whiteman et al, but the consensus in the literature is that this is not a serious risk. See this blog summary for the layman, and also Gail Whiteman, Chris Hope, and Peter Wadhams, "Climate Science: Vast Costs of Arctic Change," Comments and Opinion, Nature, July 24, 2013, https://doi.org/10.1038/499401a; E. a. G. Schuur et al., "Climate Change and the Permafrost Carbon Feedback," *Nature* 520, no. 7546 (April 2015): 171–79, https://doi.org/10.1038/nature14338.
[149] IPCC, *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, 2014), 11.
[150] Peter U. Clark et al., "Consequences of Twenty-First-Century Policy for Multi-Millennial Climate and Sea-Level Change," *Nature Climate Change* advance online publication (February 8, 2016), https://doi.org/10.1038/nclimate2923.

Agriculture

Crop models show relatively modest effects of warming of up to 5°C on yields of the major food crops, with yields increasing by 10% for some crops, and declining by 20% for others.[151] In a world of rising demand for food and ongoing poverty, reductions in yield of this kind are likely to bring major humanitarian costs. However, this would occur in the context of rising agricultural productivity: according to some estimates, yields for the major food crops are projected to grow by around 50% by 2050.[152] Thus, the effects on agriculture would not threaten to undermine the level of population that we can sustain today. In general, very cold, low-CO_2 environments are worse for agriculture than warmer environments because plants need CO_2 to grow and frost shortens the growing season.[153] Thus, while the effects on agriculture would be bad, it does not appear that they would threaten the global viability of agriculture.

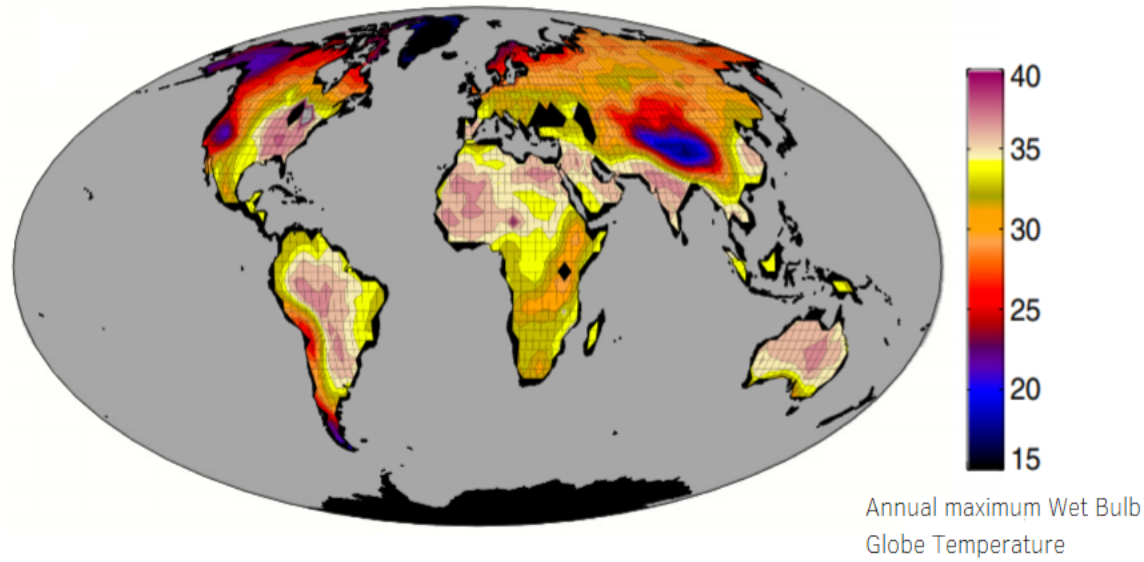
Heat stress

Extreme warming of more than 6°C would threaten the habitability of large portions of the planet, especially the tropics.[154] The Wet Bulb Globe Temperature is an indicator of heat stress. If it rises above 35°C for extended periods, people could not survive outside for long. Figure 2.7 shows the effects of warming of 12°C on Wet Bulb Globe Temperature.

[151] IPCC, *Climate Change 2014: Impacts, Adaptation, and Vulnerability: Summary for Policymakers* (Cambridge University Press, 2014), 498.
[152] Deepak K. Ray et al., "Yield Trends Are Insufficient to Double Global Crop Production by 2050," *PLOS ONE* 8, no. 6 (June 19, 2013): e66428, https://doi.org/10.1371/journal.pone.0066428.
[153] Peter J. Richerson, Robert Boyd, and Robert L. Bettinger, "Was Agriculture Impossible during the Pleistocene but Mandatory during the Holocene? A Climate Change Hypothesis," *American Antiquity* 66, no. 3 (2001): 387–411.
[154] King et al., "Climate Change–a Risk Assessment," chap. 10.

Figure 2.7.

Heat stress in a world 12°C warmer than today



Source: Sherwood and Huber, 'An adaptability limit to climate change due to heat stress', PNAS (2010)

Figure 2.7 shows that regions holding the majority of the world's population, *as people are currently distributed*, would become uninhabitable with 12°C of warming. However, 12°C of warming would, as mentioned above, take several centuries to occur, giving people lots of time to adapt by moving to higher latitudes. Thus, it does not seem that heat stress could directly cause extinction.

Biodiversity

A point of controversy in the scientific literature concerns the projected effects of extreme warming on biodiversity.[155] Theoretical models suggest substantial effects on biodiversity,[156] but the historical climatic record shows that extreme regional or global warming has not been

[155] For discussion of the disagreement, see IPCC, *Climate Change: Impacts*, 301.
[156] For discussion, see K. J. Willis and G. M. MacDonald, "Long-Term Ecological Records and Their Relevance to Climate Change Predictions for a Warmer World," *Annual Review of Ecology, Evolution, and Systematics* 42, no. 1 (2011): 267–87, https://doi.org/10.1146/annurev-ecolsys-102209-144704.

correlated with increasing species extinctions, with the exception of the Permian mass extinction, in which geological and climatic conditions were very different to today.[157] However, future warming is likely to occur in an importantly different context of habitat loss and pollution, which limit adaptability of species. Given the novelty of the situation, it is unclear what effect warming will have on biodiversity.

There is a further question of how biodiversity loss could affect human civilisation. Humans have made around 1% of species extinct, and if extinctions continue at current rates for the next few centuries, then we will eliminate more than 75% of the world's species.[158] It is unclear whether there is any abrupt threshold for global ecosystem collapse,[159] and there is a further question of how this could threaten the long-term future of humanity. Overall, the path from climate change to biodiversity loss to global catastrophe is very unclear and indirect.

Summary

The overall picture that emerges from looking at direct impacts is that extreme warming would make the Earth unrecognisable. Coastal cities would be flooded, island nations would disappear, and the tropics would become uninhabitable. But the overall evidence suggests that extreme warming would not directly cause a global catastrophe.[160]

Indirect risks of climate change

Extreme warming could also threaten a global catastrophe indirectly. The three main ways this seems possible are:

- Climate change causes political instability and conflict, leading to war involving weapons of mass destruction, such as engineered bioweapons or nuclear weapons.
- Extreme climate change leads to political instability, which in turn reduces our ability to deal with other threats.

[157] See for example Willis and MacDonald.
[158] Anthony D. Barnosky et al., "Has the Earth's Sixth Mass Extinction Already Arrived?," *Nature* 471 (March 2, 2011): 51.
[159] Johan Rockström et al., "Planetary Boundaries: Separating Fact from Fiction. A Response to Montoya et Al.," *Trends in Ecology & Evolution* 33, no. 4 (2018): 233–234.
[160] Mark Lynas arrives at a similar conclusion in Mark Lynas, *Six Degrees: Our Future on a Hotter Planet*, Updated ed. (London: Harper Perennial, 2008), chap. 6.

- Extreme climate change leads to the use of solar geoengineering, which in turn has catastrophic consequences.

We are not aware of any published comprehensive reviews of whether climate change could be an indirect global catastrophic risk. The 2014 IPCC *Impacts* report discusses the impact of climate change on international security and concludes that the impact of climate change is highly uncertain, but it is likely to be a stressor of conflict in various ways.[161]

Extreme climate change leading to nuclear war

We have seen above that extreme climate change would make drastic changes to life as we currently know it. If there were warming of more than 6°C, millions of people would probably have to relocate due to sea-level rise, extreme weather and heat stress. As of 2017, there were nearly 20 million refugees,[162] and the refugee crisis has had major political repercussions. The number of future climate refugees in the event of extreme warming could potentially exceed this by an order of magnitude and would create the biggest migration crisis in history. It might be argued that this could destabilise the political order to such an extent that the risk of nuclear war would be increased, in turn increasing the risk of civilisation-threatening nuclear winter.

If one does accept this argument, from the point of global catastrophic risk reduction, working on reducing the risk of nuclear war seems like a better option than working on climate change. The probability of nuclear winter is much greater than the probability of nuclear winter caused by climate change. The causal chain from climate change to nuclear winter is very indirect:

> Emissions => extreme climate change => mass migration => political instability => nuclear war => nuclear winter => global catastrophe

There is huge uncertainty about whether each stage in the causal chain will lead to the next. If our actions stand any chance of affecting the part of the causal chain from "political stability" to "global catastrophe", it makes more sense to focus on that. Thus, while climate change would be a stressor of this risk, from the point of view of reducing global catastrophic risk, philanthropic money would probably be better spent on other problems, on the margin.

[161] IPCC, *Climate Change: Impacts*, chap. 12.
[162] See UNHCR Popstats.

Extreme climate change undermines our ability to deal with other risks

Another possibility is that climate change leads to mass migration, which undermines our political institutions, in turn undermining our ability to deal with other global catastrophic threats, such as engineered bioweapons and AI. If this is true, as above, the path from emissions to global catastrophe is again very indirect. As above, the most cost-effective philanthropic strategy would therefore likely focus directly on those other global catastrophic risks, though society as a whole would be wise to put significant effort into reducing climate risk.

Extreme climate change leads to catastrophic use of solar geoengineering

Solar geoengineering is a form of climate engineering that involves reducing warming by reflecting sunlight back to space. The most researched form is known as *stratospheric aerosol injection* — the injection of particles, such as sulphur dioxide, into the upper atmosphere. Solar geoengineering is the only known way to quickly and relatively cheaply reduce global temperatures. However, some have argued that solar geoengineering itself introduces global catastrophic risks.[163]

The main proposed direct global catastrophic risk associated with solar geoengineering is *termination shock*. This is the worry that, due to some other catastrophe, solar geoengineering is suddenly terminated leading to rapid and highly damaging warming.[164] However, some recent work, which we find plausible, has cast doubt on the potential severity of termination shock, showing that a highly specific catastrophe would be required for it to be a threat.[165]

There is also a worry that solar geoengineering could, by some currently unknown process, cause an environmental catastrophe. However, volcanoes are natural analogues for solar geoengineering because they also cool the earth by injecting aerosols into the stratosphere. For example, the 1991 Mount Pinatubo eruption injected around 20 million tonnes of sulphur dioxide into the atmosphere, causing global temperatures to drop by half a degree.[166] If we chose to do solar geoengineering, it

[163] For an overview, see John Halstead, "Stratospheric Aerosol Injection Research and Existential Risk," *Futures*, March 9, 2018, https://doi.org/10.1016/j.futures.2018.03.004.
[164] Seth D. Baum, Timothy M. Maher, and Jacob Haqq-Misra, "Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse," *Environment Systems & Decisions* 33, no. 1 (January 8, 2013): 168–80, https://doi.org/10.1007/s10669-012-9429-y.
[165] Andy Parker and Peter J. Irvine, "The Risk of Termination Shock From Solar Geoengineering," *Earth's Future* 6, no. 3 (March 1, 2018): 456–67, https://doi.org/10.1002/2017EF000735.
[166] National Academy of Sciences, *Climate Intervention: Reflecting Sunlight to Cool Earth* (Washington, D.C.: National Academies Press, 2015), 7.

would be more controlled than some of the largest volcanic eruptions, which reduces concern about unknown effects.

There is also a concern that solar geoengineering could indirectly cause an global catastrophe by heightening political tensions. Solar geoengineering would affect the weather in all regions, and if it had real or perceived detrimental effects on some regions, those regions could blame solar geoengineering, increasing the risk of interstate conflict, which could in turn increase the risk of nuclear war or reduce our ability to manage other risks.[167] If so, the above observations again apply: focusing directly on those other risks seems likely to be more impactful on the margin (but there is still a strong case for society as a whole to reduce the risks associated with climate change and geoengineering).

Summary

The risk posed by climate change is underappreciated in many circles. The threat of more than 6°C of warming is so severe that significant concern about, and strong action against, climate change is clearly warranted.

However, the direct damage done, while extremely bad, would in our view fall short of global catastrophe. Climate change would, moreover, be a very indirect stressor of other potential global catastrophic risks. But for philanthropists aiming to make the biggest reduction to global catastrophic risk on the margin, work on other risks is likely to be a better bet, in part for the reasons outlined in the above section.

For donors who wish to support climate charities, as we discuss in our climate change report, we believe that careful philanthropists can have outsized impact by donating to our recommended climate charities.

2.5. Natural risks

Over the course of our 200,000 year history, *Homo sapiens* have avoided extinction from natural risks, which suggests that these present a fairly small risk,[168] (with the arguable exception of natural pandemics, discussed above). Indeed, it may be more relevant to consider the prospects of the *Homo* genus, which includes our own species *Homo sapiens*, as well as our ancestors, such as Neanderthals and *Homo erectus*. The *Homo* genus has survived for six million years without being

[167] Halstead, "Stratospheric Aerosol Injection Research and Existential Risk."
[168] Snyder-Beattie, Ord, and Bonsall, "An Upper Bound for the Background Rate of Human Extinction."

killed off by one of these natural risks, which suggests that the risk from these is lower still.[169] The leading natural global catastrophic risks we are currently aware of include natural pandemics (discussed above in section 2.2) supervolcanoes, Near Earth Objects (NEOs) like asteroids and comets, and gamma-ray bursts. All of these have been posited as causes of the five previous great mass extinctions.[170]

Near-Earth Objects

Many scientists believe that the dinosaurs were killed off by a large asteroid impact in Chicxulub, Mexico 65.5 million years ago.[171] This asteroid, around 10 km in diameter, would have caused earthquakes and tsunamis upon impact, and ejected huge amounts of dust, water and gas into the atmosphere, making drastic changes to the climate.[172]

NEO-tracking efforts suggest that the risk of NEOs appears small. According to a 2010 report by the US National Academy of Sciences, impacts with a diameter of 1.5 km would likely kill roughly 10% of the world population, increasing to the whole population for NEOs with a diameter of upwards of 10 km.[173] On average, NEOs with a width of ~10 km will strike Earth once every 100 million years.[174] It is thought that ~94% of nearby asteroids with a diameter of 1 km or more have been discovered, and NASA believes all asteroids with a diameter of 10 km or more have been detected.[175] Continued detection of both asteroids and comets would give us time to prepare if a large NEO were on course to hit Earth, although it is unclear whether we possess the technical capacity to deflect an NEO larger than 10 km in diameter.[176]

[169] Ord, "Will We Cause Our Own Extinction? Natural versus Anthropogenic Extinction Risks."
[170] See Arnon Dar, "Influence of Supernovae, Gamma-Ray Bursts, Solar Flares, and Cosmic Rays on the Terrestrial Environment," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (Oxford: Oxford University Press, 2008).
[171] Peter Schulte et al., "The Chicxulub Asteroid Impact and Mass Extinction at the Cretaceous-Paleogene Boundary," *Science* 327, no. 5970 (March 5, 2010): 1214–18, https://doi.org/10.1126/science.1177265.
[172] Schulte et al., 1216–17.
[173] National Research Council (U. S.). Committee to Review Near-Earth-Object Surveys and Hazard Mitigation Strategies, *Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies* (Washington, DC: National Academies Press, 2010), 23.
[174] National Research Council (U. S.). Committee to Review Near-Earth-Object Surveys and Hazard Mitigation Strategies, 19.
[175] Dr Alan Harris, personal email correspondence, 11th July 2016.
[176] National Research Council (U. S.). Committee to Review Near-Earth-Object Surveys and Hazard Mitigation Strategies, *Defending Planet Earth*, 78–79.

Supervolcanic eruptions

In 1815, the Tambora volcano erupted, killing more than 71,000 Indonesians on the islands of Lombok and Sumbawa.[177] The eruption also had global effects because it ejected large amounts of sulphur into the upper atmosphere, reflecting sunlight and causing global cooling.[178] Consequently, 1816 became known as the 'year without summer' in parts of North America and Europe: in June 1816, frosts were reported in Connecticut and snow fell in Albany, New York.[179] A much larger eruption could potentially threaten global civilisation by causing much more severe global cooling that would destroy agriculture. Indeed, some scientists have argued that the eruption of the Toba volcano in Indonesia 74,000 years ago caused a severe bottleneck in global human population,[180] though this is controversial.[181]

The magnitude of volcanic eruptions is measured by the Volcanic Explosivity Index (VEI), a logarithmic scale on which each additional point corresponds to a tenfold increase in magnitude: Tambora had a VEI of 7 and Toba a VEI of 8, so Toba was ten times more severe than Tambora.[182] Volcanic eruptions with a VEI of 8 or above are labelled 'supervolcanic eruptions'.[183] There is large uncertainty about the frequency of VEI=8 eruptions because they are very rare, meaning that we have to rely on uncertain geological proxies.[184] VEI=8 eruptions are estimated to occur on the order of every 10,000 to 100,000 years.[185] If so, it seems extremely unlikely that VEI=8 eruptions could cause an global catastrophe: humanity and our ancestors would have gone through this between

[177] Clive Oppenheimer, "Climatic, Environmental and Human Consequences of the Largest Known Historic Eruption: Tambora Volcano (Indonesia) 1815," *Progress in Physical Geography* 27, no. 2 (January 6, 2003): 230, https://doi.org/10.1191/0309133303pp379ra.
[178] Oppenheimer, 230.
[179] Oppenheimer, 244.
[180] Stanley H. Ambrose, "Did the Super-Eruption of Toba Cause a Human Population Bottleneck? Reply to Gathorne-Hardy and Harcourt-Smith," *Journal of Human Evolution* 45, no. 3 (2003): 231–237.
[181] Naomi Lubick, "Giant Eruption Cut Down to Size," Science | AAAS, November 24, 2010, https://www.sciencemag.org/news/2010/11/giant-eruption-cut-down-size.
[182] W. Aspinall et al., "Volcano Hazard and Exposure in GFDRR Priority Countries and Risk Mitigation Measures," *Volcano Risk Study 0100806-00-1-R*, 2011, 4, http://globalvolcanomodel.org/wp-content/uploads/2014/01/Aspinall_et_al_GFDRR_Volcano_Risk_Final.pdf.
[183] Michael Rampino, "Super-Volcanism and Other Geophysical Processes of Catastrophic Import," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (Oxford: Oxford University Press, 2008).
[184] Aspinall et al., "Volcano Hazard and Exposure in GFDRR Priority Countries and Risk Mitigation Measures," 30.
[185] For the lower estimate, see Jonathan Rougier et al., "The Global Magnitude–Frequency Relationship for Large Explosive Volcanic Eruptions," *Earth and Planetary Science Letters* 482 (January 15, 2018): 621–29, https://doi.org/10.1016/j.epsl.2017.11.015. For the higher estimate, see Susan Loughlin et al., *Global Volcanic Hazards and Risk* (Cambridge University Press, 2015), 97.

60 and 600 times and survived, and later flourished, at much lower levels of technological sophistication than today.

One way to reduce the risk of supervolcanic eruptions would be to improve our resilience to a 'supervolcanic winter' by developing foods that do not depend on sunlight.[186] Other methods for reducing the risk of supervolcanoes have been proposed, but have not yet been fully explored in the academic literature.[187]

Gamma-Ray Bursts

Gamma-ray bursts are narrow beams of energetic radiation probably produced by supernova explosions or mergers between highly compact objects such as neutron stars and black holes.[188] A sufficiently close, long and powerful gamma-ray burst pointed at the Earth would chiefly do damage through massive ozone depletion leading to increased UVB radiation. In addition, large amounts of nitrous oxide would be released into the atmosphere leading to reduced sunlight and global cooling.[189] Fortunately, potentially extinction-level gamma-ray bursts are extremely rare: they are estimated to occur in the order of once every one hundred million years or more.[190] Given their frequency, they might have been responsible for previous mass extinctions.[191] In principle, we may be able to predict gamma-ray bursts,[192] and the best way to prepare may again be to develop foods that do not depend on sunlight.[193]

[186] Denkenberger and Pearce, *Feeding Everyone No Matter What*.
[187] David C. Denkenberger and Robert W. Blair, "Interventions That May Prevent or Mollify Supervolcanic Eruptions," *Futures*, Futures of research in catastrophic and existential risk, 102 (September 1, 2018): 51–62, https://doi.org/10.1016/j.futures.2018.01.002.
[188] Brian C. Thomas, "Gamma-Ray Bursts as a Threat to Life on Earth," *International Journal of Astrobiology* 8, no. 3 (2009): 183–86.
[189] Thomas.
[190] See Table 2 in Tsvi Piran and Raul Jimenez, "Possible Role of Gamma Ray Bursts on Life Extinction in the Universe," *Physical Review Letters* 113, no. 23 (December 5, 2014): 231102, https://doi.org/10.1103/PhysRevLett.113.231102.
[191] A.l. Melott et al., "Did a Gamma-Ray Burst Initiate the Late Ordovician Mass Extinction?," *International Journal of Astrobiology* 3, no. 01 (January 2004): 55–61, https://doi.org/10.1017/S1473550404001910.
[192] Brian Thomas, personal correspondence, 8th July 2016
[193] Denkenberger and Pearce, *Feeding Everyone No Matter What*.