Government
Office for Science

INNOVATION: MANAGING RISK, NOT AVOIDING IT

Evidence and Case Studies

INNOVATION: MANAGING RISK, NOT AVOIDING IT

Evidence and Case Studies

This volume comprises chapters which form the evidence for the Government Chief Scientific Advisor's Annual Report 2014, together with illustrative case studies. It should be cited as:
Annual Report of the Government Chief Scientific Adviser 2014. Innovation: Managing Risk, Not Avoiding It. Evidence and Case Studies.

The Government Office for Science would like to thank the authors who contributed chapters, case studies and their time towards this report and gave it freely. A full list of authors can be found on pages 6-8.

This report is intended for:
Policy-makers, legislators, and a wide range of business people, professionals, researchers and other individuals whose interests relate to the link between risk and innovation.

The report project team was David Bennett, Graeme Collinson, Mike Edbury, Elizabeth Surkovic and Jack Wardle.

Nick Beckstead (University of Oxford) and Toby Ord (University of Oxford)

CHAPTER 10: MANAGING EXISTENTIAL RISK FROM EMERGING TECHNOLOGIES

Despite the political and organizational challenges, policymakers need to take account of low-probability, high-impact risks that could threaten the premature extinction of humanity.

Historically, the risks that have arisen from emerging technologies have been small when compared with their benefits. The potential exceptions are unprecedented risks that could threaten large parts of the globe, or even our very survival[1].

Technology has significantly improved lives in the United Kingdom and the rest of the world. Over the past 150 years, we have become much more prosperous. During this time, the UK average income rose by more than a factor of seven in real terms, much of this driven by improving technology. This increased prosperity has taken millions of people out of absolute poverty and has given everyone many more freedoms in their lives. The past 150 years also saw historically unprecedented improvements in health, with life expectancy in the United Kingdom steadily increasing by two to three years each decade. From a starting point of about 40 years, it has doubled to 80 years[2].

These improvements are not entirely due to technological advances, of course, but a large fraction of them are. We have seen the cost of goods fall dramatically due to mass production, domestic time freed up via labour saving machines at home, and people connected by automobiles, railroads, airplanes, telephones, television, and the Internet. Health has improved through widespread improvements in sanitation, vaccines, antibiotics, blood transfusions, pharmaceuticals, and surgical techniques.

These benefits significantly outweigh many kinds of risks that emerging technologies bring, such as those that could threaten workers in industry, local communities, consumers, or the environment. After all, the dramatic improvements in prosperity and health already include all the economic and health costs of accidents and inadvertent consequences during technological development and deployment, and the balance is still overwhelmingly positive.

This is not to say that governance does or should ignore mundane risks from new technologies in the future. Good governance may have substantially decreased the risks that we faced over the previous two centuries, and if through careful policy choices we can reduce future risks without much negative impact on these emerging technologies, then we certainly should do so.

However, we may not yet have seen the effects of the most important risks from technological innovation. Over the next few decades, certain technological advances may pose significant and unprecedented global risks. Advances in the biosciences and biotechnology may make it possible to create bioweapons more dangerous than any disease humanity has faced so far; geoengineering technologies could give individual countries the ability to unilaterally alter the global climate (see case study); rapid advances in artificial intelligence could give a single country a decisive strategic advantage. These scenarios are extreme, but they are recognized as potential low-probability high-impact events by relevant experts. To safely navigate these risks, and harness the potentially great benefits of these new technologies, we must continue to develop our understanding of them and ensure that the institutions responsible for monitoring them and developing policy responses are fit for purpose.

Technology has significantly improved lives in the United Kingdom and the rest of the world.

This chapter explores the high-consequence risks that we can already anticipate; explains market and political challenges to adequately managing these risks; and discusses what we can do today to ensure that we achieve the potential of these technologies while keeping catastrophic threats to an acceptably low level. We need to be on our guard to ensure we are equipped to deal with these risks, have the regulatory vocabulary to manage them appropriately, and continue to develop the adaptive institutions necessary for mounting reasonable responses.

Anthropogenic existential risks vs. natural existential risks

An *existential risk* is defined as a risk that threatens the premature extinction of humanity, or the permanent and drastic destruction of its potential for desirable future development. These risks could originate in nature (as in a large asteroid impact, gamma-ray burst, supernova, supervolcano eruption, or pandemic) or through human action (as in a nuclear war, or in other cases we discuss below). This chapter focuses on anthropogenic existential risks because — as we will now argue — the probability of these risks appears significantly greater.

Historical evidence shows that species like ours are not destroyed by natural catastrophes very often. Humans have existed for 200,000 years. Our closest ancestor, *Homo erectus*, survived for about 1.8 million years. The median mammalian species lasts for about 2.2 million years[3]. Assuming that the distribution of natural existential catastrophes has not changed, we would have been unlikely to survive as long as we have if the chance of natural extinction in a given century were greater than 1 in 500 or 1 in 5,000 (since $(1 - 1/500)^{2,000}$ and $(1 - 1/5,000)^{18,000}$ are both less than 2%). Consistent with this general argument, all natural existential risks are believed to have very small probabilities of destroying humanity in the coming century[4].

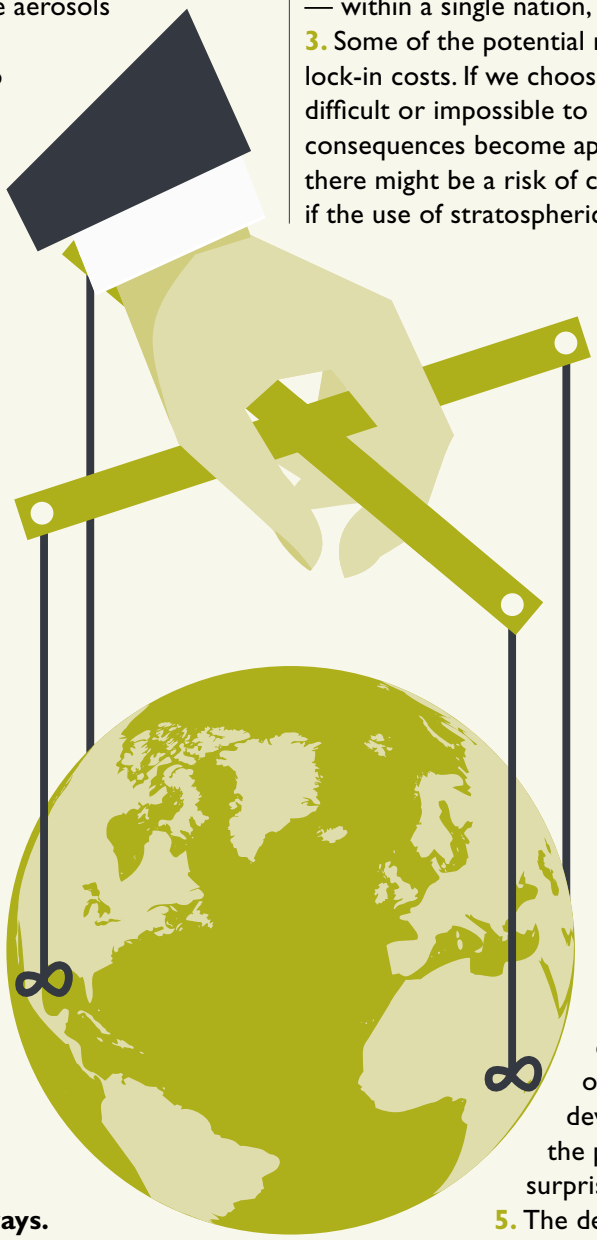
In contrast, the tentative historical evidence we do have points in the opposite direction for anthropogenic risks. The development of nuclear fission, and the atomic bomb, was the first time in history that a technology created the possibility of destroying most or all of the world's population. Fortunately we have not yet seen a global nuclear catastrophe, but we have come extremely close.

CASE STUDY

POLICY, DECISION-MAKING AND EXISTENTIAL RISK

Huw Price and Seán Ó hÉigeartaigh (University of Cambridge)

Geoengineering is the deliberate use of technology to alter planet-scale characteristics of the Earth, such as its climatic system. Geoengineering techniques have been proposed as a defence against global warming. For example, sulphate aerosols have a global cooling effect: by pumping sulphate aerosols into the high atmosphere, it may be possible to decrease global temperatures. Alternatively, seeding suitable ocean areas with comparatively small amounts of iron might increase plankton growth sufficiently to sequester significant quantities of atmospheric carbon dioxide. These technologies are already within reach, or nearly so (although their efficacy is still difficult to predict). As global warming worsens, the case for using one or more of them to ameliorate the causes or avert the effects of climate change may strengthen. Yet the long-term consequences of these techniques are poorly understood, and there may be a risk of global catastrophe if they were to be deployed, for example through unexpected effects on the global climate or the marine ecosystem.



This example illustrates the policy dimensions of existential risk in several ways.

1. It involves potentially beneficial technologies that may come with a small (though difficult to assess) risk of catastrophic side effects.

2. These risks are associated with the fact that the technology is global in impact. If we choose to employ it, we are putting all our eggs in one basket. This is especially obvious in the case of geoengineering, because the technology is intended to have planet-level effects. But it is also true of other potential sources of existential risk, such as synthetic biology or artificial intelligence, in the sense that it is unlikely that these technologies could be deployed merely locally — within a single nation, for example.

3. Some of the potential risks are associated with lock-in costs. If we choose one path now, it may be difficult or impossible to retreat later if unintended consequences become apparent — for example, there might be a risk of catastrophic sudden warming if the use of stratospheric aerosols was suddenly discontinued.

4. Once the technology is available, making a choice on its use is unavoidable — even a decision to do nothing is still a decision. Whatever we decide, our choice will have long-term consequences. However, geoengineering technology differs from some other potential sources of existential risk in that not using it is a feasible option, perhaps even the default option (at least for the time being). In other cases, various short-term benefits and associated commercial factors are likely to provide strong incentives to develop the technologies in question, and the task of managing extreme risks is to find opportunities to steer that development in order to reduce the probability of catastrophic surprises.

5. The decision to deploy geoengineering technology could, in principle, be made by a single nation or even a wealthy individual. In this respect, too, geoengineering illustrates one of the characteristic features of extreme technological risks: they are associated with the fact that powerful technologies put more power into fewer hands.

US President John F. Kennedy later confessed that during the Cuban missile crisis, the chances of a nuclear war with Russia seemed to him at the time to be "somewhere between one out of three and even". In light of this evidence, it is intuitively rather unclear that we could survive 500 or 5,000 centuries without facing a technologically-driven global catastrophe such as a nuclear war. We argue that in the coming decades, the world can expect to see several powerful new technologies that — by accident or design — may pose equal or greater risks for humanity.

1. Engineered Pathogens

Pandemics such as Spanish flu and HIV have killed tens of millions of people. Smallpox alone was responsible for more than 300 million deaths in the first half of the twentieth century. As the ongoing Ebola epidemic reminds us, disease outbreaks remain a potent threat today. However, pressures from natural selection limit the destructive potential of pathogens because a sufficiently virulent, transmissible pathogen would eliminate the host population. As others have argued, and we reiterate below, bioengineering could be used to overcome natural limits on virulence and transmissibility, allowing pandemics of unprecedented scale and severity.

For an example of an increase in fatality rates, consider mousepox, a disease that is normally non-lethal in mice. In 2001, Australian researchers modified mousepox, accidentally increasing its fatality rate to 60%, even in mice with immunity to the original version[5]. By 2003, researchers led by Mark Buller found a way to increase the fatality rate to 100%, although the team also found therapies that could protect mice from the engineered version[6].

For an example of an increase in transmissibility, consider the 'gain of function' experiments on influenza that have enabled airborne transmission of modified strains of H5N1 between ferrets[7]. Proponents of such experiments argue that further efforts building on their research "have contributed to our understanding of host adaptation by influenza viruses, the development of vaccines and therapeutics, and improved [disease] surveillance"[8]. However, opponents argue that enhancing the transmissibility of H5N1 does little to aid in vaccine development; that long lag times between capturing and sequencing natural flu samples limits the value of this work for surveillance; and that epistasis — in which interactions between genes modulate their overall effects — limits our ability to infer the likely consequences of other genetic mutations in influenza from what we have observed in gain-of-function research so far[9].

Many concerns have been expressed about the catastrophic and existential risks associated with engineered pathogens. For example, George Church, a pioneer in the field of synthetic biology, has said:

"While the likelihood of misuse of oligos to gain access to nearly extinct human viruses (e.g. polio) or novel pathogens (like IL4-poxvirus) is small, the consequences loom larger than chemical and nuclear weapons, since biohazards are inexpensive, can spread rapidly world-wide and evolve on their own."[10]

Similarly, Richard Posner[11], Nathan Myhrvold[12], and Martin Rees[13] have argued that in the future, an engineered pathogen with the appropriate combination of virulence, transmissibility and delay of onset in symptoms would pose an existential threat to humanity. Unfortunately, developments in this field will be much more challenging to control than nuclear weapons because the knowledge and equipment needed to engineer viruses is modest in comparison with what is required to create a nuclear weapon[14]. It is possible that once the field has matured over the next few decades, a single undetected terrorist group would be able to develop and deploy engineered pathogens. By the time the field is mature and its knowledge and tools are distributed across the world, it may be very challenging to defend against such a risk.

This argues for the continuing development of active policy-oriented research, an intelligence service to ensure that we know what misuse some technologies are being put to, and a mature and adaptive regulatory structure in order to ensure that civilian use of materials can be appropriately developed to maximize benefit and minimize risk.

We raise these potential risks to highlight some worst-case scenarios that deserve further consideration. Advances in these fields are likely to have significant positive consequences in medicine, energy, and agriculture. They may even play an important role in reducing the risk of pandemics, which currently pose a greater threat than the risks described here.

2. Artificial Intelligence

Artificial intelligence (AI) is the science and engineering of intelligent machines. Narrow AI systems — such as Deep Blue, stock trading algorithms, or IBM's Watson — work only in specific domains. In contrast, some researchers are working on AI with general capabilities, which aim to think and plan across all the domains that humans can. This general sort of AI only exists in very primitive forms today[15].

Many people have argued that long-term developments in artificial intelligence could have catastrophic consequences for humanity in the coming century[16], while others are more skeptical[17]. AI researchers have differing views about when AI systems with advanced general capabilities might be developed, whether such development poses significant risks, and how seriously radical scenarios should be taken. As we'll see, there are even differing views about how to characterize the distribution of opinion in the field.

In 2012, Müller and Bostrom surveyed the 100 most-cited AI researchers to ask them when advanced AI systems

> Reduction of the risk of an existential catastrophe is a global public good.

might be developed, and what the likely consequences would be. The survey defined a "high-level machine intelligence" (HLMI) as a machine "that can carry out most human professions at least as well as a typical human", and asked the researchers about which year they would assign a 10%, 50% or 90% subjective probability to such AI being developed. They also asked whether the overall consequences for humanity would be "extremely good", "on balance good", "more or less neutral", "on balance bad", or "extremely bad (existential catastrophe)".

The researchers received 29 responses: the median respondent assigned a 10% chance of HLMI by 2024, a 50% chance of HLMI by 2050, and a 90% chance of HLMI by 2070. For the impact on humanity, the median respondent assigned 20% to "extremely good", 40% to "on balance good", 19% to "more or less neutral", 13% to "on balance bad", and 8% to "extremely bad (existential catastrophe)"[18].

In our view, it would be a mistake to take these researchers' probability estimates at face value, for several reasons. First, the AI researchers' true expertise is in developing AI systems, not forecasting the consequences for society from radical developments in the field. Second, predictions about the future of AI have a mixed historical track record[19]. Third, these 'subjective probabilities' represent individuals' personal degrees of confidence, and cannot be taken to be any kind of precise estimate of an objective chance. Fourth, only 29 out of 100 researchers responded to the survey, which therefore may not be representative of the field as a whole.

The difficulty in assessing risks from AI is brought out further by a report from the Association for the Advancement of Artificial Intelligence (AAAI), which came to a different conclusion. In February 2009, about 20 leading researchers in AI met to discuss the social impacts of advances in their field. One of three sub-groups focused on potentially radical long-term implications of progress in artificial intelligence. They discussed the possibility of rapid increases in the capabilities of intelligent systems, as well as the possibility of humans losing control of machine intelligences that they had created. The overall perspective and recommendations were summarized as follows:

- "The first focus group explored concerns expressed by lay people — and as popularized in science fiction for decades — about the long-term outcomes of AI research. Panelists reviewed and assessed popular expectations and concerns. The focus group noted a tendency for the general public, science-fiction writers, and futurists to dwell on radical long-term outcomes of AI research, while overlooking the broad spectrum of opportunities and challenges with developing and fielding applications that leverage different aspects of machine intelligence."
- "There was overall skepticism about the prospect of an intelligence explosion as well as of a "coming singularity," and also about the large-scale loss of control of intelligent systems. Nevertheless, there was a shared sense that additional research would be valuable on methods for understanding and verifying the range of behaviors of complex computational systems to minimize unexpected outcomes."
- "The group suggested outreach and communication to people and organizations about the low likelihood of the radical outcomes, sharing the rationale for the overall comfort of scientists in this realm, and for the need to educate people outside the AI research community about the promise of AI for enhancing the quality of human life in numerous ways, coupled with a re-focusing of attention on actionable, shorter-term challenges."[20]

> Over the next few decades, certain technological advances may pose significant and unprecedented global risks.

This panel gathered prominent people in the field to discuss the social implications of advances in AI in response to concerns from the public and other researchers. They reported on their views about the concerns, recommended plausible avenues for deeper investigation, and highlighted the possible upsides of progress in addition to discussing the downsides. These were valuable contributions.

However, the event had shortcomings as well. First, there is reason to doubt that the AAAI panel succeeded in accurately reporting the field's level of concern about future developments in AI. Recent commentary on these issues from AI researchers has struck a different tone. For instance, the survey discussed above seems to indicate more widespread concern. Moreover, Stuart Russell — a leader in the field and author of the most-used textbook in AI — has begun publicly discussing AI as a potential existential risk[21].

In addition, the AAAI panel did not significantly engage with concerned researchers and members of the public, who had no representatives at the conference, and the AAAI panel did not explain their reasons for being sceptical of concerns about the long-term implications of AI, contrary to standard recommendations for 'inclusion' or 'engagement' in the field of responsible innovation[22]. In place of arguments, they offered language suggesting that these concerns were primarily held by "non-experts" and belonged in the realm of science fiction. It's questionable whether there is genuine expertise in predicting the long-term future of AI at all[23], and unclear how much better AI researchers would be than other informed people. But this kind of dismissal is especially questionable in light of the fact that many AI researchers in the survey mentioned above thought the risk of "extremely bad" outcomes for humanity from long-term

progress in AI had probabilities that were far from negligible. At present, there is no indication that the concerns of the public and researchers in other fields have been assuaged by the AAAI panel's interim report or any subsequent outreach effort.

What then, if anything, can we infer from these two different pieces of work? The survey suggests that some AI researchers believe that the development of advanced AI systems poses non-negligible risks of extremely bad outcomes for humanity, whilst the AAAI panel was skeptical of radical outcomes. Under these circumstances, it is impossible to rule out the possibility of a genuine risk, making a case for deeper investigation of the potential problem and the possible responses and including long-term risks from AI in horizon-scanning efforts by government.

Challenges of managing existential risks from emerging technology

Existential risks from emerging technologies pose distinctive challenges for regulation, for the following reasons:

1. The stakes involved in an existential catastrophe are extremely large, so even an extremely small risk can carry an unacceptably large expected cost[24]. Therefore, we should seek a high degree of certainty that all reasonable steps have been taken to minimize existential risks with a sufficient baseline of scientific plausibility.
2. All of the technologies discussed above are likely to be difficult to control (much harder than nuclear weapons). Small states or even non-state actors may eventually be able to cause major global problems.
3. The development of these technologies may be unexpectedly rapid, catching the political world off guard. This highlights the importance of carefully considering existential risks in the context of horizon-scanning efforts, foresight programs, risk and uncertainty assessments, and policy-oriented research.
4. Unlike risks with smaller stakes, we cannot rely on learning to manage existential risks through trial and error. Instead, it is important for government to investigate potential existential risks and develop appropriate responses even when the potential threat and options for mitigating it are highly uncertain or speculative.

As we seek to maintain and develop the adaptive institutions necessary to manage existential risks from emerging technologies, there are some political challenges that are worth considering:

1. Reduction of the risk of an existential catastrophe is a global public good, because everyone benefits[25]. Markets typically undersupply global public goods, and large-scale cooperation is often required to overcome this. Even a large country acting in the interests of its citizens may have incentives to underinvest in ameliorating existential risk. For some threats the situation may be even worse, since even a single non-compliant country could pose severe problems.
2. The measures we take to prepare for existential risks from emerging technology will inevitably be speculative, making it hard to achieve consensus about how to respond.
3. Actions we might take to ameliorate these risks are likely to involve regulation. The costs of such regulation would likely be concentrated on the regulators and the industries, whereas the benefits would be widely dispersed and largely invisible — a classic recipe for regulatory failure.
4. Many of the benefits of minimizing existential risks accrue to future generations, and their interests are inherently difficult to incorporate into political decision-making.

The stakes involved in an existential catastrophe are extremely large, so even an extremely small risk can carry an unacceptably large expected cost.

Conclusion

In the coming decades, we may face existential risks from a number of sources including the development of engineered pathogens, advanced AI, or geoengineering. In response, we must consider these potential risks in the context of horizon-scanning efforts, foresight programs, risk and uncertainty assessments, and policy-oriented research. This may involve significant political and coordination challenges, but given the high stakes we must take reasonable steps to ensure that we fully realize the potential gains from these technologies while keeping any existential risks to an absolute minimum.