Unprecedented Technological Risks

Over the next few decades, the continued development of dual-use technologies will provide major benefits to society. They will also pose significant and unprecedented global risks, including risks of new weapons of mass destruction, arms races, or the accidental deaths of billions of people. Synthetic biology, if more widely accessible, would give terrorist groups the ability to synthesise pathogens more dangerous than smallpox; geoengineering technologies would give single countries the power to dramatically alter the earth's climate; distributed manufacturing could lead to nuclear proliferation on a much wider scale; and rapid advances in artificial intelligence could give a single country a decisive strategic advantage. These scenarios might seem extreme or outlandish. But they are widely recognised as significant risks by experts in the relevant fields. To safely navigate these risks, and harness the potentially great benefits of these new technologies, we must proactively provide research, assessment, monitoring, and guidance, on a global level.

This report gives an overview of these risks and their importance, focusing on risks of extreme catastrophe, which we believe to be particularly neglected. The report explains why market and political circumstances have led to a deficit of regulation on these issues, and offers some policy proposals as starting points for how these risks could be addressed.

September 2014

Executive Summary

The development of nuclear weapons was, at the time, an *unprecedented technological risk*. The destructive power of the first atomic bomb was one thousand times greater than other weapons; and hydrogen bombs increased that destructive power one thousand-fold again. Importantly, this technological development was extremely rapid. The bombing of Hiroshima and Nagasaki came a mere six years after Einstein's initial warning letter to President Roosevelt. Nuclear technology created a significant risk of the deaths of hundreds of millions, which was openly acknowledged, with President John F. Kennedy later putting the odds of a nuclear holocaust at "somewhere between one out of three and even."

In the near future, major technological developments will give rise to new unprecedented risks. In particular, like nuclear technology, developments in synthetic biology, geoengineering, distributed manufacturing and artificial intelligence create risks of catastrophe on a global scale. These new technologies will have very large benefits to humankind. But, without proper regulation, they risk the creation of new weapons of mass destruction, the start of a new arms race, or catastrophe through accidental misuse. Some experts have suggested that these technologies are even more worrying than nuclear weapons, because they are more difficult to control. Whereas nuclear weapons require the rare and controllable resources of uranium-235 or plutonium-239, once these new technologies are developed, they will be very difficult to regulate and easily accessible to small countries or even terrorist groups.

Moreover, these risks are currently under-regulated, for a number of reasons. Protection against such risks is a global public good and thus undersupplied by the market. Implementation often requires cooperation among many governments, which adds political complexity. Due to the unprecedented nature of the risks, there is little or no previous experience from which to draw lessons and form policy. And the beneficiaries of preventative policy include people who have no sway over current political processes — our children and grandchildren.

Given the unpredictable nature of technological progress, development of these technologies may be unexpectedly rapid. A political reaction to these technologies only when they are already on the brink of development may therefore be too late. We need to implement prudent and proactive policy measures in the near future, even if no such breakthroughs currently appear imminent.

Policy to control these risks should aim at:

- Decreasing the chance of bad outcomes.
    - For example, a member country could propose to the UN that there should be guidance ensuring intergovernmental transparency and accountability on new potentially dangerous technological development.
- Improving our ability to respond if bad outcomes do occur.
    - For example, investment in early-detection monitoring for new pathogens and general-purpose vaccine, antiviral, and antibiotic development.
- Improving our current state of knowledge.
    - For example, commissioning a review to provide a detailed assessment of the risks from new technologies and to recommended policies.

Table of Contents

Nick Beckstead, Future of Humanity Institute, University of Oxford

Nick Bostrom, Director, Future of Humanity Institute, University of Oxford

Niel Bowerman, Global Priorities Project, Centre for Effective Altruism; Department of Physics, University of Oxford

Owen Cotton-Barratt, Global Priorities Project, Centre for Effective Altruism; Future of Humanity Institute, University of Oxford

William MacAskill, Uehiro Centre for Practical Ethics, University of Oxford

Seán Ó hÉigeartaigh, Cambridge Centre for the Study of Existential Risk; Future of Humanity Institute, University of Oxford

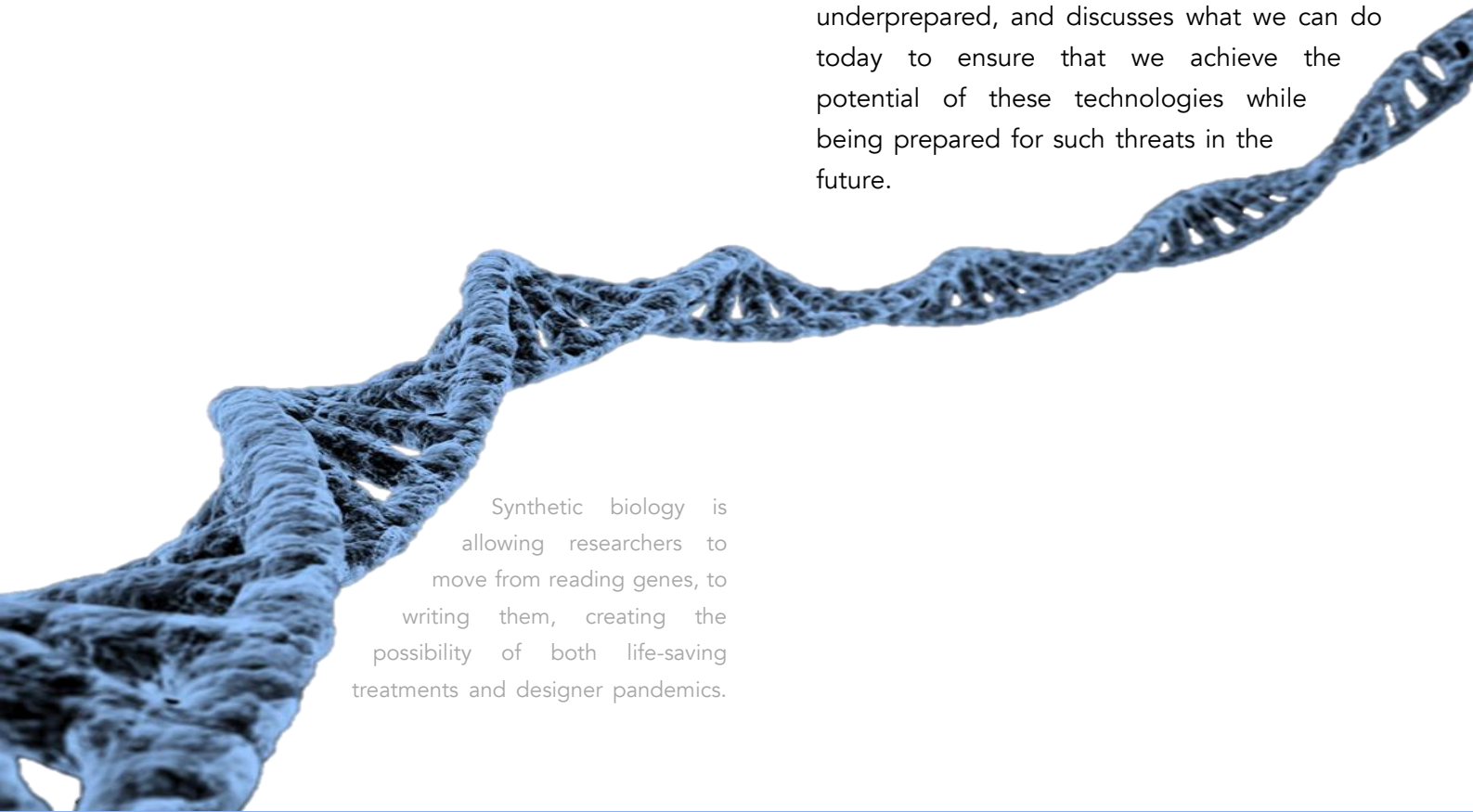
Toby Ord, Programme on the Impacts of Future Technology, Oxford Martin School, University of Oxford

Introduction

The history of civilisation is in large part a history of technological change. Many new technologies have caused large societal shifts or upset the existing geopolitical balance. Technological developments have led to vast increases in human welfare, and this trend seems set to continue. But while technological change provides very many benefits, it can also generate major new 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.

In the coming decades we can expect to see several powerful new technologies, which by accident or design may pose equal or greater risks for humanity. We have been lucky so far, but we should not trust to luck every time. This briefing explores the risks we can already anticipate, explains why we are probably underprepared, and discusses what we can do today to ensure that we achieve the potential of these technologies while being prepared for such threats in the future.



Synthetic biology is allowing researchers to move from reading genes, to writing them, creating the possibility of both life-saving treatments and designer pandemics.

The Risks

Advances in synthetic biology, geoengineering, distributed manufacturing, and artificial intelligence may all pose risks of global catastrophe. Distinguished scholars who have expressed concern about such technologies include Professor John von Neumann[1], Professor Stephen Hawking[2], Lord Professor Martin Rees[3], and the most cited legal theorist of the 20th century, US judge Richard Posner[4]. Participants at a seminal conference on global risks at Oxford University in 2008 rated the chance of a technologically induced global catastrophe during the next century at greater than ten percent[5].

Synthetic Biology

Synthetic biology is the design and construction of biological devices and systems. It has great potential as a means of developing new beneficial medical technologies. But it also creates the ability to design and build novel pathogens.

Pandemics such as Spanish Flu and smallpox have killed hundreds of millions of people in the past. However, pressures from natural selection limit the destructive potential of pathogens. Synthetic biology can overcome these natural limits, allowing pandemics of unprecedented scale. Particularly worrying is a combination of high lethality, high infectiousness, and long incubation periods: properties that can occur individually in nature, but would cause a global catastrophe if combined. Top legal theorist Judge Richard Posner vividly describes a possible worst-case scenario:

> "Religious terrorists and rogue scientists create a strain of the smallpox virus that is incurable, is immune to vaccine, and kills all its victims, rather than just 30 percent as in the case of natural smallpox. In a single round-the-world flight, a biological Unabomber, dropping off inconspicuous aerosol dispensers in major airports, infects several thousand people with the juiced-up smallpox. In the 12 to 14 days before symptoms appear, each of the initially infected victims infects five or six others, who in turn infect five or six others, and so on. Within a month more than 100 million people are infected, including almost all health workers and other "first responders," making it impossible to establish and enforce a quarantine. Before a vaccine or cure can be found, all but a few human beings, living in remote places, have died. Lacking the requisite research skills and production facilities, the remnant cannot control the disease and soon succumb as well."

This technology will be much more challenging to control than nuclear weapons because the knowledge and equipment needed to engineer viruses may be modest in comparison with what is required to create a nuclear weapon. It is plausible that once the technology is here, a single undetected terrorist group would be able to develop and deploy engineered pathogens.

Geoengineering

Geoengineering is the deliberate use of technology to alter the Earth's climatic system. Geoengineering techniques have been proposed as a last resort against global warming. For example, sulphate aerosols have a global cooling effect: by pumping sulphate aerosols into the atmosphere, it is possible to decrease global temperatures.

The technology to do this is already within reach[6]. As global warming worsens, it might become in the interests of a single country or a sub-state actor to unilaterally use geoengineering techniques in order to avert the effects of climate change. However, the consequences of these techniques are poorly understood, and there is therefore a risk of global catastrophe if they were to be deployed, through unexpected effects on

the global climate. Potentially catastrophic effects include drought, acid rain, and ozone depletion[7].

Distributed Manufacturing

Distributed manufacturing is a set of technologies that allow products to be designed and built without centralised factories. This offers the potential for extreme customisation, reduced transportation costs, and just-in-time production, but also bypasses government controls on manufactured goods such as destructive weapons. The rapid growth of 3D printing is an early demonstration of the economic value of this technology, and has already generated security risks by allowing people to create functional homemade firearms[8].

An extreme of this trend is atomically precise manufacturing via small, home-based nanotechnological fabricators. If achieved, these would allow the distributed manufacturing of a very wide variety of products for very modest costs. Development of this technology could make arms control far more difficult. Another key threat is giving a first-mover country a rapid increase in economic power, including the power to develop new weapons en masse, leading to geopolitical instability and potentially a global arms race.

Artificial General Intelligence

Artificial Intelligence (AI) is the science and engineering of creating intelligent machines. Narrow AI systems — such as chess playing algorithms, stock trading algorithms, or IBM's Watson — work only in specific domains. In contrast, some researchers are working on Artificial General Intelligence (AGI), which aims to think and plan across all the domains that humans can, rather than just in specific areas. AGI only exists in very primitive forms today. However, the computing power of the leading supercomputers now comes close to that of the human brain, and a survey conducted in 2012 found that the leading AI researchers believe that there is a 10% chance that within two decades researchers will develop AGI systems capable of doing most jobs that humans do today, rising to a 50% probability of such systems by 2050[9].

AGI would give advantages across a broad range of areas: in computational resources, communication speed, serial depth of thought, duplicability and editability, goal coordination, and rationality. For that reason, there could be swift progress from roughly human-level AGI to AGI systems that vastly outstrip human capabilities. Such a rapid shift could concentrate power in the hands of a single group or nation. If some actors control highly advanced artificial intelligence while others do not, they could gain a decisive strategic advantage over all others. General intelligence underlies human capabilities in strategizing, social manipulation, hacking, technology research, and economic productivity. An AGI system with a sufficiently strong advantage along any of these dimensions could mean a decisive advantage for its controllers.

There are also significant risks from accidents during development. Even the primitive AIs we have today have been known to behave in highly unpredictable ways in order to complete their tasks. Similarly, above-human-level AGIs might take unexpected and highly destructive actions if these happened to lie on some unforeseen path to completing the task set to them.

Leading AI researchers believed that, though very good outcomes for humanity were the most likely consequence, there is a 7% chance of an "extremely bad" outcome or "existential catastrophe" from developing advanced AGI systems[10].

Assessing the Risks

Unprecedented technological risks such as those above raise distinctive concerns:

- They pose the potential for loss of life on a massive scale: hundreds of millions or billions of deaths; in some cases, even threatening the survival of the human race.
- They may be much more difficult to control than nuclear weapons. Small states or even non-state actors may be able to cause major global problems.
- The development of these technologies may be unexpectedly rapid, catching the political world off-guard.
- We cannot use historical frequencies to gauge their probability of occurring, so our judgements will necessarily include some degree of speculation.
- We cannot rely on learning how to handle them through experience or by trial and error.

Evaluating Catastrophic Risks

Though these technologies have the potential for a wide array of benefits, the bad outcomes, if they do occur, are truly catastrophic.

The deployment of new weapons of mass destruction could kill hundreds of millions or billions of people. Even small probabilities of this outcome would give an expected loss on the order of millions of deaths.

As well as sheer loss of life, such a bad outcome could lead to the collapse of modern civilisation in many or all parts of the world, undoing centuries of progress[10]. Even from a narrowly economic perspective, the cost of a tiny risk of such a global catastrophe is staggering.

Even assuming no long-term economic growth, and a 3% discount rate, today's gross world product of $70 trillion per year implies that the present value of all future economic activity exceeds $2,000 trillion. If the Oxford-based risk estimate of 10% is distributed evenly over the century, and we conservatively assume a global catastrophe would wipe out one tenth of the value of civilization, then the expected value of insuring against technological catastrophe is at least $200 billion each year. And this figure does not account for risk aversion, nor include the non-financial worth of humanity's scientific, artistic, and moral progress.

Finally, some of these technologies, in particular synthetic biology and AGI, pose risks to the continued survival of the human race. The importance of preventing this from happening is difficult to overestimate. Only a small fraction of the people who may ever live have already been born, and human extinction therefore represents a loss of potential on a tremendous scale[11]. As the physicist Carl Sagan puts it:

> "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…. 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 ten million years), we are talking about some 500 trillion people yet to come."[12]

Market and Political Complexity

If these risks are significant, and if addressing them is so important, one might wonder why there hasn't already been more effort to reduce them. However, several factors suggest that we should expect a large market failure in the area.

First, reduction of the risk of a global catastrophe is a global public good, as everyone benefits and it is hard even for a large country doing such work to capture more than a small proportion of the value. Markets typically

undersupply such goods, and large-scale cooperation is required to overcome this[13]. For some threats the situation may be even worse, since even a single non-compliant country could pose severe problems.

Second, they are unprecedented. Risks tend to capture attention after an event of that type transpires. But the nature of these risks may make learning from experience impossible. So the measures we take against such risks are inevitably speculative. In a normal market, speculators can use their judgement to assess unprecedented scenarios, so these risks are priced according to aggregate best judgement by the market even if only a few people are considering them. For global public goods, however, cooperation is needed, so all parties must come to an agreement on how to value the risk. This is a much taller order.

Third, actions we might take to ameliorate these risks are likely to involve regulation. We can expect additional market failure here: the costs are concentrated (on the regulators and the industries), whereas the benefits are widely dispersed. When safety processes are functioning these benefits may be largely invisible. This is a typical case where lobbying is one-sided and so may hold too much sway.

Finally, any global catastrophe threatens to affect not just the people of today but the people of tomorrow — our children and grandchildren. Many of the benefits of minimising such risks accrue to them: it is therefore not just a public good, but an intergenerational public good. Since future generations have no say on decisions made today, it is likely their interests are under-represented in the status quo. Moreover, short election cycles mean that political actors have limited incentive to deal with these longer-term considerations.

Policy Proposals

Though these risks emerge from a diverse array of technologies, they can be evaluated and assessed in similar ways, and prudent policy responses are often similar. From a policy perspective, it therefore makes sense to address the class of unprecedented risks as a whole.

Many of the risks are still too unknown and in some cases too far off for us to be confident in policies that act against a specific outcome. Therefore policy response today should make sure we are well-placed to act swiftly and appropriately in the future. These policies can be divided into two categories: (i) improving our state of knowledge; (ii) building safety into our institutions.

Learning More

The issue of unprecedented technological risk is complex and deserves further research. To learn more about this topic, a national government or intergovernmental agency could:

- Fund scientists and engineers in key areas to research and report on possible risks arising from their field, including exactly what circumstances would lead to bad outcomes, and sensitivity analysis on their assumptions.
- Include unprecedented technological risks in horizon-scanning projects and risk registers with appropriate timelines.
- Commission an independent review of unprecedented risks, for example on the model of the Stern Review on the Economics of Climate Change, produced for the British government.

All of these would help by improving our state of knowledge about exactly which outcomes we should protect against, and what can be done to avoid them. This would lead to better targeted policies in the future, and reduce the chance of being taken by surprise by a new technology.

Establishing safe governance and culture

We should also strive to build systems that will avoid the market failures common to this area, enabling us to act swiftly and appropriately in response to emerging risks. Policies of this type we could adopt include:

- When creating or updating governance structures, include explicit pathways for accountability to the rights and needs of future generations. This would help to mitigate against the unduly short-term focus in decision-making.
- Foster an active culture of safety in relevant areas, similar to the nuclear safety culture. This would make safety a collective goal, and reduce market failures from misaligned incentives.
- Create a government point of contact to allow scientists and engineers to anonymously report safety concerns. This would be similar to existing anti-terrorism hotlines, but cover potential accidents and side effects as well as malicious acts, and would have the expertise necessary to respond quickly and appropriately.
- Require research institutes in potentially dangerous areas to internalise some of the costs associated with the risks of their research by requiring them to have insurance against catastrophic accident. This would both incentivise research in safer areas when they have similar upside, and encourage the development of better safety protocols.

Longer term

Longer term, we will want to introduce policies which mitigate the risks, or reduce the bad effects if they occur. The threats primarily come from two sources: accidents and malicious intent. By limiting research, making research safer, and preventing dangerous research from reaching malicious actors, we can lower the chances of threats occurring. If a bad outcome does occur, then we can improve our response by: increasing the time available to respond to the threat via better forecasting and detection; improving the tools available to respond; and improving the coordination and execution of the response.

Some examples of policies that a national government could implement are as follows:

- Give a government body the oversight of public funding of research in the highest risk areas. A recent example of dangerous research, in need of regulation, was the development of an airborne variant of avian flu by Dutch researchers[14].
- Require all researchers in particularly dangerous areas to register on a central database; this would make it harder for terrorists to operate in the area undetected.
- Require DNA synthesis companies to:
    - Use commercially available software to screen all incoming orders for toxicity and infectivity.
    - Verify customer details, and maintain records of all customers and orders in case there is a suspected bioterror threat.

- Set up an initiative to give developing countries access to safe technologies in exchange for setting up safety and monitoring systems to protect against accidents and terrorism. This was proposed for biotechnology by former Assistant Secretary General to the UN, Professor Stephen Stedman. It aims to imitate the success of nuclear non-proliferation initiatives.
- Collaborate with or imitate the US IARPA ACE program for predicting future global events, and include forecasting of long-term technological trends and disasters[15]. By using subsidized real money prediction markets and other mechanisms for knowledge creation, aggregation, and elicitation, this would give access to expert-level judgements that would update swiftly. This in turn would give a longer lead-time to allow response to possibly imminent threats.
- Fund the development of broad-spectrum vaccines, antivirals and antibiotics that could quickly be adapted to work against new diseases, both natural and engineered.
- Subsidise the development of safe virtual environments for AI development and testing, so that new intelligences are by default tested within such an environment.
- Develop early-warning systems to detect bio-threats, particularly at ports.
- Develop national and international emergency response plans, focusing on the most extreme scenarios, to ensure society can continue to function while key decision-makers may be infected.

The proposals we list here are offered as an informed starting point — demonstrating the types of concrete action that could help to mitigate these threats. There remains room for policy and domain experts to revise, extend, and supplement these ideas to better address the risks.

---

[1] Von Neumann, J. (1958). The Computer and the Brain, (New Haven: Yale University Press).
[2] Hawking S. (2010), "Abandon Earth or Face Extinction", Bigthink.com, 6 August 2010.
[3] Rees, M. (2003). *Our final century* (p. 42). London: Heinemann.
[4] Posner Richard, A. (2004). Catastrophe: Risk and Response.
[5] Sandberg, A. & Bostrom, N. (2008): "Global Catastrophic Risks Survey", Technical Report #2008-1, Future of Humanity Institute, Oxford University: pp. 1-5.
[6] Robock, A., A. Marquardt, B. Kravitz, and G. Stenchikov (2009), Benefits, risks, and costs of stratospheric geoengineering, Geophys. Res. Lett., 36, L19703, doi:10.1029/2009GL039209
[7] Heckendorn, P; Weisenstein, D; Fueglistaler, S; Luo, B P; Rozanov, E; Schraner, M; Thomason, L W; Peter, T (2009). "The impact of geoengineering aerosols on stratospheric temperature and ozone". *Environmental Research Letters* **4** (4): 045108.
[8] http://www.bbc.co.uk/news/science-environment-22421185
[9] Sandberg, A. & Bostrom, N. (2011): "Machine Intelligence Survey", Technical Report #2011-1, Future of Humanity Institute, Oxford University: pp. 1-12.
[10] Bostrom, N., & Cirkovic, M. M. (Eds.). (2011). *Global catastrophic risks*. Oxford University Press.
[11] Matheny, J. G. (2007). Reducing the risk of human extinction. *Risk analysis*,*27*(5), 1335-1344.
[12] Sagan, Carl (1983). "Nuclear war and climatic catastrophe: Some policy implications". *Foreign Affairs* **62**: 275.
[13] Kaul, Inge, Isabelle Grunberg and Marc A. Stern (eds.) (1999). *Global public goods: international cooperation in the 21st century*. NY: Oxford University Press
[14] http://www.nytimes.com/2012/06/22/health/h5n1-bird-flu-research-that-stoked-fears-is-published.html?_r=0
[15] http://www.iarpa.gov/Programs/ia/ACE/ace.html

The Global Priorities Project brings new analysis to the problem of how to allocate scarce resources between diverse global priorities such as education, health, enterprise, and future generations.

The Future of Humanity Institute is a multidisciplinary research institute at the University of Oxford. It enables a select set of leading intellects to bring the tools of mathematics, philosophy, and science to bear on big-picture questions about humanity and its prospects.

The Oxford Martin Programme on the Impacts of Future Technology analyses possibilities related to long-range technological change and the potential social impacts of future transformative technologies.

The Centre for the Study of Existential Risk is a multidisciplinary research centre at the University of Cambridge dedicated to the study and mitigation of risks that could lead to human extinction.