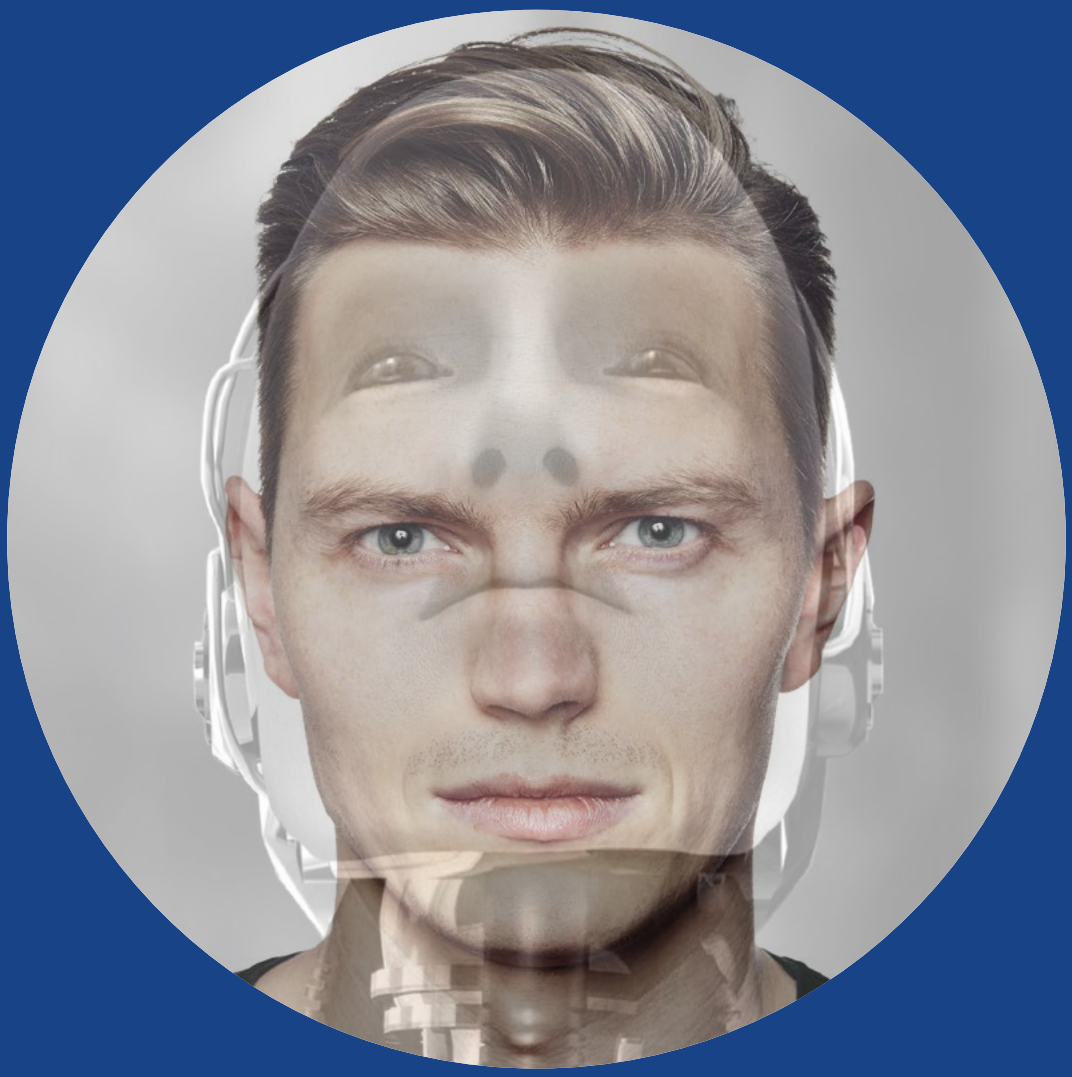

The Next Step

EXPONENTIAL LIFE

Technological Wild Cards: Existential Risk and a Changing Humanity

SEÁN Ó HÉIGEARTAIGH

Opening image:
Chris Jordan
Crushed Cars #2, Tacoma (2004)
"Intolerable Beauty: Portraits of American Mass Consumption" Series
44 × 62 cm.

Seán Ó hÉigeartaigh
Centre for the Study of Existential Risk (CSER), Cambridge, UK

Seán Ó hÉigeartaigh is the Executive Director of Cambridge's Centre for the Study of Existential Risk (CSER), a world-leading academic research center focused on global risks and emerging technologies. He is codeveloper of the Centre for the Future of Intelligence, a new Cambridge-Oxford-Imperial-Berkeley collaboration on the opportunities and challenges of artificial intelligence. Prior to Cambridge, he ran interdisciplinary research programs on emerging technologies, and on catastrophic risk modelling, at Oxford's Future of Humanity Institute. His research interests include: emerging technologies, risk, technology policy, horizon-scanning, and foresight. He has a PhD in Genomics from Trinity College Dublin.

MORE ABOUT THE AUTHOR [+]

Humanity has always faced threats to its global survival, such as asteroid impacts and supervolcanoes. Yet now the greatest risks we face may be a result of our own scientific and civilizational progress. We are developing technologies of unprecedented power, such as nuclear weapons and engineered organisms. We are also wiping out species, changing the climate, and burning through the earth's resources at an unsustainable rate as the global population soars. However, the coming century's breakthroughs in science and technology will also provide powerful solutions to many of the greatest challenges we face.

A NEW ERA OF RISK

In the early hours of September 26, 1983, Stanislav Petrov was on duty at a secret bunker outside Moscow. A lieutenant colonel in the Soviet Air Defense Forces, his job was to monitor the Soviet early warning system for nuclear attack. Tensions were high; earlier that month Soviet jets had shot down a Korean civilian airliner, an event US President Reagan had called " a crime against humanity that must never be forgotten." The KGB had sent out a flash message to its operatives to prepare for possible nuclear war.

Petrov's system reported a US missile launch. He remained calm, suspecting a computer error. The system then reported a second, third, fourth, and fifth launch. Alarms screamed, and lights flashed. Petrov "had a funny feeling in my gut"[1]; why would the US start a nuclear war with only five missiles? Without any additional evidence available, he radioed in a false alarm.

Later, it emerged that sunlight glinting off clouds at an unusual angle had triggered the system.

This was not an isolated incident. Humanity came to the brink of large-scale nuclear war many times during the Cold War.[2] Sometimes computer system failures were to blame, and human intuition saved the day. Sometimes human judgment was to blame, but cooler heads prevented thermonuclear war. Sometimes flocks of geese were enough to trigger the system. As late as 1995, a Norwegian weather rocket launch resulted in the nuclear briefcase being open in front of Russia's President Yeltsin.

If each of these events represented a coin flip, in which a slightly different circumstance—a different officer in a different place, in a different frame of mind—could have resulted in nuclear war, then we have played some frightening odds in the last seventy-odd years. And we have been breathtakingly fortunate.

Humanity has already changed a lot over its lifetime as a species. While our biology is not drastically different than it was 70,000 years ago, the capabilities enabled by our scientific, technological, and sociocultural achievements have changed what it is to be human. Whether through the processes of agriculture, the invention of the steam engine, or the practices of storing and passing on knowledge and ideas, and working together effectively as large groups, we have dramatically augmented our biological abilities. We can lift heavier things than our biology allows, store and access more information than our brains can hold, and collectively solve problems that we could not individually.

The species will change even more over coming decades and centuries, as we develop the ability to modify our biology, extend our abilities through various forms of human-machine interaction, and continue the process of sociocultural innovation. The long-term future holds tremendous promise: continued progress may allow humanity to spread throughout a galaxy that to the best of our knowledge appears devoid of intelligent life. However, what we will be in the future may bear little resemblance to what we are now, both physically and in terms of capability. Our descendants may be augmented far beyond what we currently recognize as human.

This is reflected in the careful wording of Nick Bostrom's definition of existential risk, the standard definition used in the field. An existential risk "is one that threatens the premature extinction of earth-originating intelligent life, or the permanent and drastic destruction of its potential for desirable future development."[3] Scholars in the field are less concerned about the form humanity may take in the long-term future, and more concerned that we avoid circumstances that might prevent our descendants—whatever form they may take—from having the opportunity to flourish.

One way in which this could happen is if a cataclysmic event were to wipe out our species (and perhaps, with it, the capacity for our planet to bear intelligent life in future). But another way would be if a cataclysm fell short of human extinction, but changed our circumstances such that further progress became impossible. For

example, runaway climate change might not eliminate all of us, but might leave so few of us, scattered at the poles, and so limited in terms of accessible resources, that further scientific, technological, and cultural progress might become impossible. Instead of spreading to the stars, we might remain locked in a perennial battle for survival in a much less bountiful world.

The Risks We Have Always Faced

For the first 200,000 years of humanity's history, the risks that have threatened our species as a whole have remained relatively constant. Indonesia's crater lake Toba is the result of a catastrophic volcanic super-eruption that occurred 75,000 years ago, blasting an estimated 2800 cubic kilometers of material into the atmosphere. An erupted mass just 1/100th of this from the Tambora eruption (the largest in recent history) was enough to cause the 1816 "year without a summer," where interference with crop yields caused mass food shortages across the northern hemisphere. Some lines of evidence suggest that the Toba event may have wiped out a large majority of the human population at the time, although this is debated. At the Chixculub Crater in Mexico, geologists uncovered the scars of the meteor that most likely wiped out seventy-five percent of species on earth at that time, including the dinosaurs, sixty-six million years ago. This may have opened the door, in terms of available niches, for the emergence of mammalian species and ultimately humanity.

Reaching further into the earth's history uncovers other, even more cataclysmic events for previous species. The Permian-Triassic extinction event wiped out 90–96% of species at the time. Possible causes include meteor impacts, rapid climate change possibly due to increased methane release, large-scale volcanic activity, or a combination of these. Even further back, the cyanobacteria that introduced oxygen to our atmosphere, and paved the way for oxygen-breathing life, did so at a cost: they brought about the extinction of nearly all life at the time, to whom oxygen was poisonous, and triggered a "snowball earth" ice age.

The threats posed by meteor or asteroid impacts and supervolcanoes have not gone away. In principle an asteroid could hit us at any point with little warning. A number of geological hotspots could trigger a volcanic eruption; most famously, the Yellowstone Hotspot is believed to be "due" for another massive explosive eruption.

However, on the timescale of human civilization, these risks are very unlikely in the coming century, or indeed any given century. 660,000 centuries have passed since the event that wiped out the dinosaurs; the chances that the next such event will happen in our lifetimes is likely to be of the order of one in a million. And "due around now" for Yellowstone means that geologists expect such an event at some point in the next 20,000–40,000 years. Furthermore, these threats are static; there is little evidence that their probabilities, characteristics, or modes of impact are changing significantly on a human civilizational timescale.

Yellowstone National Park (Wyoming, USA) is home to one of the planet's hot spots, where a massive volcanic explosion could someday occur.

New Challenges

New challenges have emerged alongside our civilizational progress. As we organized ourselves into larger groups and cities, it became easier for disease to spread among us. During the Middle Ages the Black Death outbreaks wiped out 30–60% of Europe's population. And our travel across the globe allowed us to bring diseases with us to places they would never have otherwise reached; following European colonization of the Americas, disease outbreaks wiped out up to 95% native populations.

The Industrial Revolution allowed huge changes in our capabilities as a species. It allowed rapid progress in scientific knowledge, engineering, and manufacturing capability. It allowed us to draw heavily from cheap, powerful, and rapidly available energy sources—fossil fuels. It helped us to support a much greater global population. The global population more than doubled between 1700 and 1850, and population in England—birthplace of the Industrial Revolution—increased from 5 to 15 million in the same period, and doubled again to 30 million by 1900.[4] In effect, these new technological capabilities allowed us to extract more resources, create much greater changes to our environment, and support more of us than had ever previously been possible. This is a path we have been accelerating along ever since then, with greater globalization, further scientific and technological development, a rising global population, and, in developed nations at least, a rising quality of life and resource use footprint.

On July 16, 1945, the day of the Trinity atomic bomb test, another milestone was reached. Humans had developed a weapon that could plausibly change the global environment in such an extreme way as to threaten the continued existence of the human species.

Power, Coordination, and Complexity

Humanity now has a far greater power to shape its environment, locally and globally, than any species that has existed to our knowledge; more so even than the cyanobacteria that turned this into a planet of oxygen-breathing life. We have repurposed huge swathes of the world's land to our purposes—as fields to produce food for us, cities to house billions of us, roads to ease our transport, mines to provide our material resources, and landfill to house our waste. We have developed structures and tools such as air conditioning and heating that allow us to populate nearly every habitat on earth, the supply networks needed to maintain us across these locations, scientific breakthroughs such as antibiotics, and practices such as sanitation and pest control to defend ourselves from the pathogens and pests of our environments. We also modify ourselves to be better adapted to our environments, for example through the use of vaccines.

This increased power over ourselves and our environment, combined with methods to network and coordinate our activities over large numbers and wide areas, has created great resilience against many threats we face. In most of the developed world we can guarantee adequate food and water access for the large majority of the population, given normal fluctuations in yield; our food sources are varied in type and geographical

location, and many countries maintain food stockpiles. Similarly, electricity grids provide a stable source of energy for developed populations, given normal fluctuations in supply. We have adequate hygiene systems and access to medical services, given normal fluctuations in disease burden, and so forth. Furthermore, we have sufficient societal stability and resources that we can support many brilliant people to work on solutions to emerging problems, or to advance our sciences and technologies to give us ever-greater tools to shape our environments, increase our quality of life, and solve our future problems.

It goes without saying that these privileges exist to a far lesser degree in developing nations, and that many of these privileges depend on often exploitative relationships with developing nations, but this is outside the scope of this chapter. Here the focus is on the resilience or vulnerability of the human-as-species, which is tied more closely to the resilience of the best-off than the vulnerability of the poorest, except to the extent that catastrophes affecting the world's most vulnerable populations would certainly impact the resilience of less vulnerable populations.

Many of the tools, networks, and processes that make us more resilient and efficient in "normal" circumstances, however, may make us more vulnerable in the face of extreme circumstances. While a moderate disruption (for example, a reduced local crop yield) can be absorbed by a network, and compensated for, a catastrophic disruption may overwhelm the entire system, and cascade into linked systems in unpredictable ways. Systems critical for human flourishing, such as food, energy, and water, are inextricably interlinked (the "food-water-energy nexus") and a disruption in one is near-guaranteed to impact the stability of the others. Further, these affect and are affected by many other human- and human-affected processes: our physical, communications, and electronic infrastructure, political stability (wars tend to both precede and follow famines), financial systems, and extreme weather (increasingly a human-affected phenomenon). These interactions are very dynamic and difficult to predict. Should the water supply from the Himalayas dry up one year, we have very little idea of the full extent of the regional and global impact, although we could reasonably speculate about droughts, major crop failures, and mass starvation, financial crises, a massive immigration crisis, regional warfare that could go nuclear and escalate internationally, and so forth. Although unlikely, it is not outside the bounds of imagination that through a series of unfortunate events, a catastrophe might escalate to one that would threaten the collapse of global civilization.

Two factors stand out.

Firstly, the processes underpinning our planet's health are interlinked in all sorts of complex ways, and our activities are serving to increase the level of complexity, interlinkage, and unpredictability—particularly in the case of extreme events.

Secondly, the fact is that, despite our various coordinated processes, we as a species are very limited in our ability to act as a globally coordinated entity, capable of taking the most rational actions in the interests of the whole—or in the best interests of our continued survival and flourishing.

This second factor manifests itself in global inequality, which benefits developed nations in some ways, but also introduces major global vulnerabilities; the droughts, famines, floods, and mass displacement of populations likely to result from the impacts of climate change in the developing world are sure to negatively affect even the richest nations. It manifests itself in an inability to act optimally in the face of many of our biggest challenges. More effective coordination on action, communication, and resource distribution would make us more resilient in the face of pandemic outbreaks, as illustrated so vividly by the Ebola outbreak of 2014; a relatively mild outbreak of what should be an easily controllable disease served to highlight how inadequate pandemic preparedness and response was.[5, 6] We were lucky that the disease was not one with greater pandemic potential, such as one capable of airborne transmission and with long incubation times.

Our limited ability to coordinate in our long-term interest manifests itself in a difficulty in limiting our global resource use, limiting the impact of our collective activities on our global habitat, and of investing our resources optimally for our long-term survival and well-being. And it limits our ability to guarantee that advances in science and technology be applied to furthering our well-being and resilience, as opposed to being destabilizing or even used for catastrophically hostile purposes, such as in the case of nuclear weapons.

Collective action problems are as old as humanity,[7] and we have made significant progress in designing effective institutions, particularly in the aftermath of World War I and II. However, the stakes related to these problems become far greater as our power to influence our environment grows—through sheer force of numbers and distribution across the planet, and through more powerful scientific and technological tools with which to achieve our myriad aims or to frustrate those of our fellows. We are entering an era in which our greatest risks are overwhelmingly likely to be caused by our own activities, and our own lack of capacity to collectively steer and limit our power.

OUR FOOTPRINT ON THE EARTH

Population and Resource Use

The United Nations estimated the earth's population at 7.4 billion as of March 2016, up from 6.1 billion in 2000, 2.5 billion in 1950, and 1.6 billion in 1900. Long-term growth is difficult to predict (being affected by many uncertain variables such as social norms, disease, and the occurrence of catastrophes) and thus varies widely between studies. However, UN projections currently point to a steady increase through the twenty-first century, albeit at a slower growth rate, reaching just shy of 11 billion in 2100.[8] Most estimates indicate global population will

eventually peak and then fall, although the point at which this will happen is very uncertain. Current estimates of resource use footprints indicate that the global population is using fifty percent more resources per year than the planet can replenish. This is likely to continue rising sharply; more quickly than the overall population. If the average person used as many resources as the average American, some estimates indicate the global population would be using resources at four times the rate that they can be replenished. The vast majority of the population does not use food, energy, and water, nor release CO_2 at the rate of the average American. However, the rapid rise of a large middle class in China is beginning to result in much greater resource use and CO_2 output in this region, and the same phenomenon is projected to occur a little later on in India.

Catastrophic Climate Change

Without a significant change of course on CO_2 emissions, the world is on course for significant human-driven global warming; according to the latest IPCC report, an increase of 2.5 to 7.8 °C can be expected under "business as usual" assumptions. The lower end of this scale will have significant negative repercussions for developing nations in particular but is unlikely to constitute a global catastrophe; however, the upper end of the scale would certainly have global catastrophic consequences. The wide range in part reflects significant uncertainty over how robust the climate system will be to the "forcing" effect of our activities. In particular, scientists focused on catastrophic climate change worry about a myriad of possible feedback loops. For example, a reduction of snow cover, which reflects the sun's heat, could increase the rate of warming resulting in greater loss of snow cover. The loss of arctic permafrost might result in the release of large amounts of methane in the atmosphere, which would accelerate the greenhouse effect further. The extent to which oceans can continue to act as both "heat sinks" and "carbon sinks" as we push the concentration of CO_2 in the atmosphere upward is unknown. Scientists theorize the existence of "tipping points," which, once reached, might trigger an irreversible shift—for example, the collapse of the West Antarctic ice sheets or the melt of Greenland's huge glaciers, or the collapse of the capacity for oceans to absorb heat and sequester CO_2. In effect, beyond a certain point, a "rollercoaster" process may be triggered, where 3 degrees of temperature rise rapidly and irreversibly may lead to 4 degrees, and then 5.

Laudable progress has been made on achieving global coordination around the goal of reducing global carbon emissions, most notably in the aftermath of the December 2015 United Nations Climate Change Conference. 174 countries signed an agreement to reach zero net anthropogenic greenhouse gas emissions by the second half of the twenty-first century, and to "pursue efforts to limit" the temperature increase to 1.5 °C. But many experts hold that these goals are unrealistic, and that the commitments and actions being taken fall far short of what will be needed. According to the International Energy Agency's Executive Director Fatih Birol: "We think we are lagging behind strongly in key technologies, and in the absence of a strong government push, those technologies will never be deployed into energy markets, and the chances of reaching the two-degree goal are very slim."[9]

Soil Erosion

Soil erosion is a natural process. However human activity has increased the global rate dramatically, with deforestation, drought, and climate change accelerating the rate of loss of

fertile soil. There are reasons to expect this trend to accelerate; some of the most powerful drivers of soil erosion are extreme weather events, and these events are expected to increase dramatically in frequency and severity as a result of climate change.

Biodiversity Loss

The world is entering an era of dramatic species extinction driven by human activity.[10] Since 1900, vertebrate species have been disappearing at more than 100 times the rate seen in non-extinction periods. In addition to the intrinsic value of the diversity of forms of life on earth (the only life-inhabited planet currently known to exist in the universe), catastrophic risk scholars worry about the consequences for human societies. Ecosystem resilience is a tremendously complex phenomenon, and it seems plausible that tipping points exist in them. For example, the collapse of one or more keystone species underpinning the stability of an ecosystem could result in a broader ecosystem collapse with potentially devastating consequences for human system stability (for example, should key pollinator species disappear, the consequences for agriculture could be profound). Current human flourishing relies heavily on these ecosystem services, but we are threatening them at an unprecedented rate, and we have a poor ability to predict the consequences of our activity.

Everything Affects Everything Else

Once again, the sheer complexity and interconnectedness of these risks represents a key challenge. None of these processes happen in isolation, and developments in one affect the others. Climate change affects ecosystems by forcing species migration (for those that can), a change in plant and animal patterns of growth and behavior, and by driving species extinction. Reductions in available soil force us to drive more deeply into nonagricultural wilderness to provide the arable land we need to feed our populations. And the ecosystems we threaten play important roles in maintaining a stable climate and environment. Recognizing that we cannot get all the answers we need on these issues by studying them in isolation, threats posed by the interplay of these phenomena are a key area of study for catastrophic risk scholars.

All these developments result in a world with greater uncertainty, the emergence of huge and unpredictable new vulnerabilities, and more extreme and unprecedented events. These events will play out in a crowded world that contains more powerful technologies, and more powerful weapons, than have ever existed before.

HUMANITY AND TECHNOLOGY IN THE TWENTY-FIRST CENTURY

Our progress in science and technology, and related civilizational advances, have allowed us to house far more people on this planet, and have provided the power for those people to influence their environment more than any previous species. This progress is not of itself a bad thing, nor is the size of our global population.

There are good reasons to think that with careful planning, this planet should be able to house seven billion or more people stably and comfortably.[11] With sustainable agricultural practices and innovative use of irrigation methods, it should be possible for many relatively

uninhabited and agriculturally unproductive parts of the world to support more people and food production. An endless population growth on a finite planet is not possible without a collapse; however, growth until the point of collapse is by no means inevitable. Stabilization of population size is strongly correlated with several factors we are making steady global progress on: including education (especially of women), and rights and a greater level of control for women over their own lives. While there are conflicting studies,[12] many experts hold that decreasing child mortality, while leading to population increase in the near-term, leads to a drop in population growth in the longer term. In other words, as we move toward a better world, we will bring about a more stable world, provided intermediate stages in this process do not trigger a collapse or lasting global harm.[13, 14]

Current advances in science and technology, while not sufficient in themselves, will play a key role in making a more resilient and sustainable future possible. Rapid progress is happening in carbon-zero energy sources such as solar photovoltaics and other renewables.[15] Energy storage remains a problem, but progress is occurring on battery efficiency. Advances in irrigation techniques and desalination technologies may allow us to provide water to areas where this has not previously been possible, allowing both food production and other processes that depend on reliable access to clean water. Advances in materials technology will have wide-ranging benefits, from lighter, more energy-efficient vehicles, to more efficient buildings and energy grids, to more powerful scientific tools and novel technological innovations. Advances in our understanding of the genetics of plants are leading to crops with greater yields, greater resilience to temperature shifts, droughts and other extreme weather, and greater resistance to pests—resulting in a reduction of the need for polluting pesticides. We are likely to see many further innovations in food production; for example, exciting advances in lab-grown meat may result in the production of meat with a fraction of the environmental footprint of livestock farming.

Many of the processes that have resulted in our current unsustainable trajectories can be traced back to the Industrial Revolution, and our widespread adoption of fossil fuels. However, the Industrial Revolution and fossil fuels must also be recognized as having unlocked a level of prosperity, and a rate and scale of scientific and technological progress that would simply not have been possible without them. While a continued reliance on fossil fuels would be catastrophic for our environment, it is unclear whether many of the "clean technology" breakthroughs that will allow us to break our dependence on fossil fuels would have been possible without the scientific breakthroughs that were enabled directly, or indirectly, by this rich, abundant, and easily available fuel source. The goal is clear: having benefitted so

tremendously from this "dirty" stage of technology, we now need to take advantage of the opportunity it gives us to move onto cleaner and more powerful next-generation energy and manufacturing technologies. The challenge will be to do so before thresholds of irreversible global consequence have been passed.

The broader challenge is that humanity as a species needs to transition to a stage of technological development and global cooperation where as a species we are "living within our means": producing and using energy, water, food, and other resources at a sustainable rate, and by methods that will not impose long-term negative consequences on our global habitat—for at least as long as we are bound to it. There are no physical reasons to think that we might not be capable of developing an extensive space-faring civilization at a future point. And if we last that long, it is likely we will develop extensive abilities to terraform extraterrestrial environments to be hospitable to us—or indeed, transform ourselves to be suitable to currently inhospitable environments. However, at present, in Martin Rees's words, there is no place in our Solar System nearly as hospitable as the most hostile environment on earth, and so we are bound to this fragile blue planet.

Part of this broader challenge is gaining a better understanding of the complex consequences of our actions, and more so, of the limits of our current understanding. Even if we cannot know everything, recognizing when our uncertainty may lead us into dangerous territory can help us figure out an appropriately cautious set of "safe operating parameters" (to borrow a phrase from Steffen et al.'s "Planetary Boundaries"[16]) for our activities. The second part of the challenge, perhaps harder still, is developing the level of global coordination and cooperation needed to stay within these safe operating parameters.

Technological Wild Cards

While much of the Centre for the Study of Existential Risk's research focuses on these challenges—climate change, ecological risks, resource use, and population, and the interaction between these—the other half of our work is on another class of factors: transformative emerging and future technologies. We might consider these "wild cards"; technological developments significant enough to change the course of human civilization significantly in and of themselves. Nuclear weapons are such a wild card; their development changed the nature of geopolitics instantly and irreversibly. They also changed the nature of

With 537 square meters of solar panels and six blocks of lithium-ion batteries, *PlanetSolar* is the world's largest solar ship, as well as its fastest. It is also the first to have sailed round the world using exclusively solar power.

global risk: now many of the stressors we worry about might escalate quite quickly through human activity to a worst-case scenario involving a large-scale exchange of nuclear missiles. The scenario of most concern from an existential risk standpoint is one that might trigger a nuclear winter: a level of destruction sufficient to send huge amounts of particulate matter into the atmosphere and cause a lengthy period of global darkness and cold. If such a period persisted for long enough, this would collapse global food production and could drive the human species to near- or full-extinction. There is disagreement among experts about the scale of nuclear exchange needed to trigger a nuclear winter, but it appears eminently plausible that the world's remaining arsenals, if launched, might be sufficient.

Nuclear weapons could be considered a wild card in a different sense: the underlying science is one that enabled the development of nuclear power, a viable carbon-zero alternative to fossil fuels. This *dual-use* characteristic—that the underlying science and technology could be applied to both destructive purposes, and peaceful ones—is common to many of the emerging technologies that we are most interested in.

A few key sciences and technologies of focus for scholars in this field include, among others:

> Topics within bioscience and bioengineering such as the manipulation and modification of certain viruses and bacteria, and the creation of organisms with novel characteristics and capabilities (genetic engineering and synthetic biology).
>
> Geoengineering: a suite of proposed large-scale technological interventions that would aim to "engineer" our climate in an effort to slow or even reverse the most severe impacts of climate change.
>
> Advances in artificial intelligence—in particular, those that relate to progress toward artificial *general* intelligence—AI systems capable of matching or surpassing human intellectual abilities across a broad range of domains and challenges.

Progress on these sciences are driven in great part by a recognition of their potential for improving our quality of life, or the role they could play in aiding us to combat existing or emerging global challenges. However, in and of themselves they may also pose large risks.

Virus Research

Despite advances in hygiene, vaccines, and other health technology, natural pandemic outbreaks remain among the most potent global threats we face; for example, the 1918 Spanish influenza outbreak killed more people than World War I. This threat is of particular concern in our increasingly crowded, interconnected world. Advances in virology research are likely to play a central role in better defenses against, and responses to, viruses with pandemic potential.

A particularly controversial area of research is "gain-of-function" virology research, which aims to modify existing viruses to give them different host transmissibility and other characteristics. Researchers engaged in such research may help identify strains with high

pandemic potential, and develop vaccines and antiviral treatment. However, research with infectious agents runs the risk of accidental release from research facilities. There have been suspected releases of infectious agents from laboratory facilities. The 1977–78 Russian influenza outbreak is strongly suspected to have originated due to a laboratory release event,[17] and in the UK, the 2007 foot-and-mouth outbreak may have originated in the Pirbright animal disease research facility.[18] Research on live infectious agents is typically done in facilities with the highest biosafety containment procedures, but some experts maintain that the potential for release, while low, remains, and may outweigh the benefits in some cases.

Some worry that advances in some of the same underlying sciences may make the development of novel, targeted biological weapons more feasible. In 2001 a research group in Australia inadvertently engineered a variant of mousepox with high lethality to vaccinated mice.[19] An accidental or deliberate release of a similarly modified virus infecting humans, or a species we depend heavily on, could have catastrophic consequences.

Similarly, *synthetic biology* may lead to a wide range of tremendous scientific benefits. The field aims to design and construct new biological parts, devices, and systems, and to comprehensively redesign living organisms to perform functions useful to us. This may result in synthetic bacterial and plant "microfactories," designed to produce new medicines, materials, and fuels, to break down waste, to act as sensors, and much more. In principle, such biofactories could be designed with much greater precision than current genetic modification and biolytic approaches. They should also allow products to be produced cheaply and cleanly. Such advances would be transformative on many challenges we currently face, such as global health care, energy, and fabrication.

Moreover, as the tools and facilities needed to engage in the science of synthetic biology become cheaper, a growing "citizen science" community is emerging around synthetic biology. Community "DIY Bio" facilities allow people to engage in novel experiments and art projects; some hobbyists even engage in synthetic biology projects in their own homes. Many of the leaders in the field are committed to synthetic biology being as open and accessible as possible worldwide, with scientific tools and expertise available freely. Competitions such as iGEM (International Genetically Engineered Machine) encourage undergraduate student teams to build and test biological systems in living cells, often with a focus on applying the science to important real-world challenges, and also to archive their results and products so as to make them available to future teams to build on.

Such citizen science represents a wonderful way of making cutting-edge science accessible and exciting to generations of

innovators. However, the increasing ease of access to increasingly powerful tools is a cause of concern to the risk community. Even if the vast majority engaging in synthetic biology are both responsible and well intentioned, the possibility of bad actors or unintended consequences (such as the release of an organism with unintended ecological consequences) exists. Further, we may expect that the range and severity of negative consequences will increase, as well as the difficulty in tracking those who have access to the necessary tools and expertise. At present, biosafety and biosecurity is deeply embedded within the major synthetic biology initiatives. In the United States, the FBI works closely with synthetic biology centers, and leaders in the field espouse the need for good practices at every level. However, this area will progress rapidly, and a balance will need to be struck between allowing access to powerful tools to a wide number of people who can do good with them, while restricting the potential for accidents or deliberate misuse. It remains to be seen how easy it will be to achieve this.

Geoengineering represents a host of challenges. Stratospheric aerosol geoengineering represents a particularly powerful proposal: here, a steady stream of reflective aerosols would be released into the upper atmosphere in order to reduce the amount of the sun's light reaching the earth's surface globally. This effectively mimics the global cooling phenomenon that occurs after a large volcanic eruption, when particulate matter is blasted into the atmosphere. However, current work is focused on theoretical modelling, with very minimal practical field tests carried out to date. Questions remain about how practically feasible it would be to achieve this on a global scale, and what impact it would have on rainfall patterns and crop growth.

It should be highlighted that this is not a solution to climate change. While global temperature might be stabilized or lowered, unless this was accompanied by reduction of CO_2 emissions, then a host of damages such as ocean acidification would still occur. Furthermore, if CO_2 emissions were allowed to continue to rise during this period, then a major risk termed "termination shock" could manifest. In this case, if any circumstance resulted in an abrupt cessation of stratospheric aerosol geoengineering, then the increased CO_2 concentration in the atmosphere would result in a rapid jump in global temperature, which would have far more severe impacts on ecosystems and human societies than the already disastrous effects of a gradual rise.

Critics fear that such research might be misunderstood as a way of avoiding the far more costly process of eliminating carbon emissions; and some are concerned that intervening

in such a profound way in our planet's functioning is deeply irresponsible. It also raises knotty questions about global governance: should any one country have the right to engage in geoengineering, and, if not, how could a globally coordinated decision be reached, particularly if different nations have different exposures to the impacts of climate change, and different levels of concern about geoengineering, given we are all under the same sky?

Proponents highlight that we may already be committed to severe global impacts from climate change at this stage, and that such techniques may allow us the necessary breathing room needed to transition to zero-carbon technology while temporarily mitigating the worst of the harms. Furthermore, unless research is carried out to assess the feasibility and likely impacts of this approach, we will not be well placed to make an informed decision at a future date, when the impacts of climate change may necessitate extreme measures. Eli Kintisch, a writer at *Science*, has famously called geoengineering "a bad idea whose time has come."[20]

***Artificial intelligence*, explored in detail in Stuart Russell's chapter, may represent the wildest card of all. Everything we have achieved in terms of our civilizational progress, and shaping the world around us to our purposes, has been a product of our intelligence. However, some of the intellectual challenges we face in the twenty-first century are ones that human intelligence alone is not best suited to: for example, sifting through and identifying patterns in huge amounts of data, and integrating information from vast and interlinked systems. From analyzing disparate sources of climate data, to millions of human genomes, to running thousands of simulations, artificial intelligence will aid our ability to make use of the huge amount of knowledge we can gather and generate, and will help us make sense of our increasingly complex, interconnected world. Already, AI is being used to optimize energy use across Google's servers, replicate intricate physics experiments, and discover new mathematical proofs. Many specific tasks traditionally requiring human intelligence, from language translation to driving on busy roads, are now becoming automatable; allowing greater efficiency and productivity, and freeing up human intelligence for the tasks that AI still cannot do. However, many of the same advances have more worrying applications; for example, allowing collection and deep analysis of data on us as individuals, and paving the road for the development of cheap, powerful, and easily scalable autonomous weapons for the battlefield.**

These advances are already having a dramatic impact on our world. However, the vast majority of these systems can be described as "narrow" AI. They can perform functions at human level or above in narrow, well-specified domains, but lack the general cognitive abilities that humans, dogs, or even rats have: general problem-solving ability in a "real-world" setting, an ability to learn from experience and apply knowledge to new situations, and so forth.

There is renewed enthusiasm for the challenge of achieving "general" AI, or AGI, which would be able to perform at human level or above across the range of environments and cognitive challenges that humans can. However, it is currently unknown how far we are from such a scientific breakthrough, or how difficult the fundamental challenges to achieving this will be, and expert opinion varies widely. Our only proof of principle is the human brain, and it will take decades of progress before we can meaningfully understand the brain to a degree that would allow us to replicate its key functions. However, if and when such a breakthrough is achieved, there is reason to think that progress from human-level general intelligence to superintelligent AGI might be achieved quite rapidly.

Improvements in the hardware and software components of AI, and related sciences and technologies, might be made rapidly with the aid of advanced general AI. It is even conceivable that AI systems might directly engage in high-level AI research, in effect accelerating the process by allowing cycles of self-improvement. A growing number of experts in AI are concerned that such a process might quickly result in extremely powerful systems beyond human control; Stuart Russell has drawn a comparison with nuclear chain reaction.

Superintelligent AI has the potential to unlock unprecedented progress on science, technology, and global challenges; to paraphrase the founders of Google DeepMind, if intelligence can be "solved," it can then be used to help solve everything else. However, the risk from this hypothetical technology, whether through deliberate use or unintended runaway consequences, could be greater than that of any technology in human history. If it is plausible that this technology might be achieved in this century, then a great deal of research and planning—both on the technical design of such systems, and the governance structures around their development—will be needed in the decades beforehand in order to achieve a desirable transition.

Predicting the Future

The field also engages in exploratory and foresight-based work on more forward-looking topics; these include future advances in neuroscience and nanotechnology, future physics experiments, and proposed manufacturing technologies that may be developed in coming decades, such as molecular manufacturing. While we are limited in what we can say in detail about future scientific breakthroughs, it is often possible to establish some useful groundwork. For example, we can identify developments that should, in principle, be possible based on our current understanding of the relevant science. And we can dismiss ideas that are pure "science fiction," or sufficiently unfeasible to be safely ignored for now, or that represent a level of progress that makes them unlikely to be achieved for many generations.

By focusing further on those that could plausibly be developed within the next half century, we can give considerations to their underlying characteristics and possible impacts on the world, and of the broad principles we might bear in mind for their safe development and application. While it would have been a fool's errand to try to predict the full impacts of

the Internet prior to 1960, or of the development of nuclear weapons prior to 1945, it would certainly be possible to develop some thinking around the possible implications of very sophisticated global communications and information-sharing networks, or of a weapon of tremendous destructive potential.

Lastly, if we have some ideas about the directions from which transformative developments might come, we can engage in foresight and road-mapping research. This can help identify otherwise insignificant breakthroughs and developments that may indicate meaningful progress toward a more transformative technology being reached, or a threshold beyond which global dynamics are likely to shift significantly (such as photovoltaics and energy storage becoming cheaper and more easily accessible than fossil fuels).

Confronting the Limits of Our Knowledge

A common theme across these emerging technologies and emerging risks is that a tremendous level of scientific uncertainty and expert disagreement typically exists. This is particularly the case for future scientific progress and capabilities, the ways in which advances in one domain may influence progress in others, and the likely global impacts and risks of projected advances. Active topics of research at CSER include how to obtain useful information from a range of experts with differing views, and how to make meaningful scientific progress on challenges where we have discontinuous data, or few case studies to draw on, or even when we must characterize an entirely unprecedented event. This might be a hypothesized ecological tipping point, which when passed would result in an irreversible march toward the collapse of an entire critical ecosystem. Or it might be a transformative scientific breakthrough such as the development of artificial general intelligence, where we only have current trends in AI capability, hardware, and expert views on the key unsolved problems in the field to draw insight from. It is unrealistic to expect that we can always, or even for the most part, be right. We need to have humility, to expect false positives, and to be able to identify priority research targets from among many weak signals.

Recognizing that there are limits to the level of detail and certainty that can be achieved, this work is often combined with work on general principles of scientific and technological governance. For example, work under the heading of "responsible innovation" focuses on the challenge of developing collective stewardship of progress in science and technology in the present, with a view to achieving good future outcomes.[21] This combines scientific foresight with processes to involve the key stakeholders at the appropriate stages of a technology's development. At different stages these stakeholders will include: scientists involved in fundamental research and applied research;

industry leaders; researchers working on the risks, benefits, and other impacts of a technology; funders; policymakers; regulators; NGOs and focus groups; and laypeople who will use or be affected by the development of a technology. In the case of technologies with a potential role in global catastrophic risk, the entire global population holds a stake. Therefore decisions with long-term consequences must not rest solely with a small group of people, represent only the values of a small subset of people, or fail to account for the likely impacts on the global population.

There have been a number of very encouraging specific examples of such foresight and collaboration, where scientific domain specialists, interdisciplinary experts, funders, and others have worked together to try to guide an emerging technology's development, establish ethical norms and safety practices, and explore its potential uses and misuses in a scientifically rigorous way. In bioengineering, the famous 1975 Asilomar conference on recombinant DNA established important precedents, and more recently summits have been held on advances such as human gene editing. In artificial intelligence, a number of important conferences have been held recently, with enthusiastic participation from academic and industry research leaders in AI alongside interdisciplinary experts and policymakers. A number of the world's leading AI research teams have established ethical advisory panels to inform and guide their scientific practices, and a cross-industry "partnership on AI to benefit people and society" involving five companies leading fundamental research has recently been announced.[22]

More broadly, it is crucial that we learn from the lessons of past technologies and, where possible, develop principles and methodologies that we can take forward. This may give us an advantage in preparing for developments that are currently beyond our horizon and that methodologies too deeply tied to specific technologies and risks may not allow. One of the key concerns associated with risks from emerging and future technologies is the rate at which progress occurs and at which the associated threats may arise. While every science will throw up specific challenges and require domain-specific techniques and expertise, any tools or methodologies that help us to intervene reliably earlier are to be welcomed. There may be a

“People were always getting ready for tomorrow. I didn’t believe in that. Tomorrow wasn’t getting ready for them. It didn’t even know they were there.”

The Road, a novel with which US writer Cormac McCarthy won the Pulitzer Prize in 2007, was later made into a movie of the same name by John Hillcoat with a script adapted by Joe Penhall.

The Road (2009), John Hillcoat.

limited window of opportunity for averting such risks. Indeed, this window may occur in the early stages of developing a technology, well before the fully mature technology is out in the world, where it is difficult to control. Once Pandora's box is open, it is very difficult to close.

WORKING ON THE (DOOMSDAY) CLOCK

Technological progress now offers us a vision of a remarkable future. The advances that have brought us onto an unsustainable pathway have also raised the quality of life dramatically for many, and have unlocked scientific directions that can lead us to a safer, cleaner, more sustainable world. With the right developments and applications of technology, in concert with advances in social, democratic, and distributional processes globally, progress can be made on all of the challenges discussed here. Advances in renewable energy and related technologies, and more efficient energy use—advances that are likely to be accelerated by progress in technologies such as artificial intelligence—can bring us to a point of zero-carbon emissions. New manufacturing capabilities provided by synthetic biology may provide cleaner ways of producing products and degrading waste. A greater scientific understanding of our natural world and the ecosystem services on which we rely will aid us in plotting a trajectory whereby critical environmental systems are maintained while allowing human flourishing. Even advances in education and women's rights globally, which will play a role in achieving a stable global population, can be aided specifically by the information, coordination, and education tools that technology provides, and more generally by growing prosperity in the relevant parts of the world.

There are catastrophic and existential risks that we will simply not be able to overcome without advances in science and technology. These include possible pandemic outbreaks, whether natural or engineered. The early identification of incoming asteroids, and approaches to shift their path, is a topic of active research at NASA and elsewhere. While currently there are no known techniques to prevent or mitigate a supervolcanic eruption, this may not be the case with the tools at our disposal a century from now. And in the longer run, a civilization that has spread permanently beyond the earth, enabled by advances in spaceflight, manufacturing, robotics, and terraforming, is one that is much more likely to endure. However, the breathtaking power of the tools we are developing is not to be taken lightly. We have been very lucky to muddle through the advent of nuclear weapons without a global catastrophe. And within this century, it

is realistic to expect that we will be able to rewrite much of biology to our purposes, intervene deliberately and in a large-scale way in the workings of our global climate, and even develop agents with intelligence that is fundamentally alien to ours, and may vastly surpass our own in some or even most domains—a development that would have uniquely unpredictable consequences.

It is reassuring to note that there are relatively few individual events that could cause an existential catastrophe—one resulting in extinction or a permanent civilizational collapse. Setting aside the very rare events (such as supervolcanoes and asteroids), the most plausible candidates include nuclear winter, extreme global warming or cooling scenarios, the accidental or deliberate release of an organism that radically altered the planet's functioning, or the release of an engineered pathogen. They also include more speculative future advances: new types of weaponry, runaway artificial intelligence, or maybe physics experiments beyond what we can currently envisage. Many global risks are, in isolation, survivable—at least for some of us—and it is likely that human civilization could recover from them in the long run: less severe global warming, various environmental disasters and ecosystem collapses, widespread starvation, most pandemic outbreaks, conventional warfare (even global).

However, this latter class of risks, and factors that might drive them (such as population, resource use, and climate change) should not be ignored in the broader study of existential risk. Nor does it make sense to consider these challenges in isolation: in our interconnected world they all affect each other. The threat of global nuclear war has not gone away, and many scholars believe that it may be rising again (at the time of writing, North Korea has just undergone its most ambitious nuclear test to date). If climate pressures, drought, famine, and other resource pressures serve to escalate geopolitical tensions, or if the potential use of a new technology, such as geoengineering, could lead to a nuclear standoff, then the result is an existential threat.

For all these reasons and more, a growing community of scholars across the world believe that the twenty-first century will see greater change and greater challenges than any century in humanity's past history. It will be a century of unprecedented global pressures, and a century in which extreme and unpredictable events are likely to happen more frequently than ever before in the past. It will also be a century in which the power of technologies unlike any

we have had in our past history will hang over us like multiple Damocles' swords. But it will also be a century in which the technologies we develop, and the institutional structures we develop, may aid us in solving many of the problems we currently face—if we guide their development, and their uses and applications, carefully.

It will be a century in which we as a species will need to learn to cooperate on a scale and depth that we have never done before, both to avoid the possibility of conflict with the weapons of cataclysmic power we have developed, and to avoid the harmful consequences of our combined activities on the planet. And despite how close we came to falling on the first hurdle with nuclear weapons, there are reasons for great optimism. The threat they presented has, indeed, led to a greater level of international cooperation, and international structures to avoid the possibility of large-scale war, than has ever happened before. In a world without nuclear weapons, we may, indeed, have seen a third world war by now. And the precedent set by international efforts around climate change mitigation is a powerful one. In December 2015, nations around the world agreed to take significant steps to reduce the likelihood of global catastrophic harm to future generations, even though in many cases these steps may be both against the individual economic interests of these nations, and against the economic interests of the current generation. With each of these steps, we learn more, and we put another plank in the scientific and institutional scaffolding we will need to respond effectively to the challenges to come.

If we get it right in this century, humanity will have a long future on earth and among the stars.

ACKNOWLEDGMENTS

Seán Ó hÉigeartaigh's work is supported by a grant from Templeton World Charity Foundation. The opinions expressed in this chapter are those of the author and do not necessarily reflect the views of Templeton World Charity Foundation.

NOTES

1. M. Garber, "The man who saved the world by doing absolutely nothing," *The Atlantic* (2013). http://www.theatlantic.com/technology/archive/2013/09/the-man-who-saved-the-world-by-doing-absolutely-nothing/280050.
2. P. M. Lewis et al., *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy* (London: Chatham House, the Royal Institute of International Affairs, 2014).
3. Nick Bostrom, "Existential risks," *Journal of Evolution and Technology* 9(1) (2002).
4. J. Jefferies, "The UK population: past, present and future," in *Focus on People and Migration* (London: Palgrave Macmillan UK, 2005).
5. B. Gates, "The next epidemic—lessons from Ebola," *New England Journal of Medicine* 372(15) (2015).
6. Jeremy J. Farrar and Peter Piot, "The Ebola emergency—immediate action, ongoing strategy," *New England Journal of Medicine* 371 (2014): 16.
7. See http://slatestarcodex.com/2014/07/30/meditations-on-moloch.
8. World Population Prospects: the 2015 Revision. United Nations Department of Economic and Social Affairs. https://esa.un.org/unpd/wpp/Publications/Files/Key_Findings_WPP_2015.pdf.
9. See https://www.technologyreview.com/s/601601/six-months-after-paris-accord-were-losing-the-climate-change-battle.
10. G. Ceballos et al., "Accelerated modern human-induced species losses: Entering the sixth mass extinction," *Science Advances* 1(5) (2015).
11. Toby Ord, "Overpopulation or underpopulation," in *Is the Planet Full?*, Ian Goldin (ed.), (Oxford: OUP, 2014).
12. J. D. Shelton, "Taking exception. Reduced mortality leads to population growth: an inconvenient truth," *Global Health: Science and Practice* 2(2) (2014).
13. See https://www.givingwhatwecan.org/post/2015/09/development-population-growth-and-mortality-fertility-link.
14. Bill Gates, "2014 Gates annual letter: 3 myths that block progress for the poor," *Gates Foundation* 14 (2014). http://www.gatesfoundation.org/Who-We-Are/Resources-and-Media/Annual-Letters-List/Annual-Letter-2014.
15. D. King et al., *A Global Apollo Programme to Combat Climate Change* (London: London School of Economics, 2015). http://www.globalapolloprogram.org/
16. W. Steffen et al., "Planetary boundaries: Guiding human development on a changing planet," *Science* 347(6223) (2015)..
17. M. Rozo and G. K. Gronvall, "The reemergent 1977 H1N1 strain and the gain-of-function debate," *mBio* 6(4) (2015).
18. T. Hugh Pennington, "Biosecurity 101: Pirbright's lessons in laboratory security," *BioSocieties* 2(04) (2007).
19. M. J. Selgelid et al., "The mousepox experience," *EMBO reports* 11(1) (2010): 18–24.
20. See https://www.wired.com/2010/03/hacktheplanet-qa.
21. J. Stilgoe et al., "Developing a framework for responsible innovation," *Research Policy* 42(9) (2013).
22. See http://www.partnershiponai.org.

OpenMind

OPENMIND CHANNEL

READ THE FULL BOOK

The Next Step: exponential life

[+]

RELATED ARTICLES

Innovation: it is Generally Agreed that Science Shapes Technology, but is that the Whole Story?

[+]

Futures Studies: Theories and Methods

[+]

Provably Beneficial Artificial Intelligence

[+]

OTHER BOOKS