# THE LONG VIEW

FIRST

# THE
# LONG VIEW

## ESSAYS ON POLICY, PHILANTHROPY, AND THE LONG-TERM FUTURE

---

EDITED BY
NATALIE CARGILL AND TYLER M. JOHN

FIRST

Published in the United Kingdom by
FIRST Strategic Insight Ltd., Victory House, 99-101 Regent Street, London W1B 4EZ



ISBN: 978-0-9957281-8-9

Published in the United Kingdom



Longview Philanthropy is an expert-led philanthropic advisory nonprofit for major donors who want to do the most good possible with their giving. We focus on grants to protect future generations, and prioritise civilisation-threatening disasters, fostering a society with a more long-term outlook, and seeding research into what our highest priorities should be. Everything we offer is free-of-charge, independent, and reviewed by external expertise.

Natalie and Tyler are grateful to Will Fenning for editorial assistance.

Victory House, 99-101 Regent Street, London, W1B 4EZ, United Kingdom
Tel: +44 20 7440 3500 Email: publisher@firstforum.org

firstforum.org

"Together, these contributions embody our heartfelt conviction that philanthropy need not relegate itself to an ameliorative practice of mere alms-giving. Every moment our entire future is at stake, and today's philanthropists possess the power, the opportunity, and the moral obligation to protect posterity and maximise the flourishing of every generation yet to come."

**Natalie Cargill and Tyler M. John, "Matter that Matters"**

# CONTENTS

# CONTENTS (CONTINUED)

# FOREWORD

by
Toby Ord
*Future of Humanity Institute, Faculty of Philosophy, University of Oxford*

Our world today is remarkable. We have a level of health, prosperity, freedom, and education of which our forebears could only dream. And we have the ability to do things of which they could not even dream – exploring the surfaces of other planets with robotic servants; or searching through archives larger than the library of Alexandria while waiting in a queue for lunch.

Yet we did not create this remarkable world. We stand on the shoulders of the 100 billion people who came before us: the ten thousand generations of humanity who each inherited their culture, institutions, and knowledge from their parents, made their own small improvements, and passed this legacy on to their children. Ten thousand links in an unbroken chain.

Individuals, groups, and nations made choices that echoed across the generations. Choices like whether to limit the power of kings, to pursue industrialisation, to end the slave trade, or to recognise the rights of women. Some choices helped us progress more quickly; some held us back; and some changed the very direction in which we were headed. Indeed some choices, such as the development of atomic weapons, have threatened the very existence of a future at all. And many of these choices were made with little regard for the people of the long-term future – the distant generations whose entire ways of life they would shape.

We too make choices that will shape the future – over decades, centuries, or millennia. We have every reason to believe that the effects of our choices upon the people of the future will be just as profound as those which shaped our own time. And we owe it to the people of the future

to make these choices with care and thought; to take their interests as seriously as we take our own. Especially on issues like climate change, where one needs no great powers of prediction to see that most of the people who stand to win or lose do not yet exist.

The essays in this book are some of our first steps towards an understanding of how to make today's choices in ways that take the people of tomorrow seriously. This is not an easy undertaking. It requires the space to look beyond the news cycle, the election cycle, the business cycle, and to see the bigger picture. It requires new ways of thinking about policy, politics, and even political systems. It requires adjustments to our economic tools for evaluating future outcomes, and to our ways of forecasting long-term effects and trends.

But we need to rise to these challenges. And if these early attempts are a taste of what is to come, I believe we shall.

# INTRODUCTION

## Matter that Matters

by
Natalie Cargill
*Founder and CEO, Longview Philanthropy*
and
Tyler M. John
*Head of Research, Longview Philanthropy, and PhD Candidate in Philosophy, Rutgers University – New Brunswick*

The most remarkable thing that ever happened occurred between 445 and 541 million years ago, during a period that is known, in geological time scale, as the Cambrian period: matter became aware. By this time period, a rocky composite of gas and stardust had been visited by interloping asteroids carrying one unusually stable, drinkable molecule that would conspire together with inorganic elements to form cells – membrane-bound structures that fight to maintain an equilibrium with their environment in an effort to survive. As these cells achieved sufficient complexity and began to intermarry, their intimate unions created organisms capable of forming representations of their environment. The birth of complex organisms brought with it awareness, goals, and the capacity for matter to organise itself towards these goals. From lifeless gas, rock, and water: purpose.

According to modern science, the pale blue dot is the only region in all of the observable universe that is home to this kind of matter. And for all of the thirteen billion years that preceded the emergence of awareness on earth, and the billion trillion stars that make up the observable universe, our planet appears to be the only region of space that has ever developed a sense of purpose.

Cosmologists tell us that our universe is still in its infancy. The last stars

will be born in over a trillion years' time. But the coals of the universe will remain warm for a million times as long, due to a steady stream of brown dwarfs fusing to keep the cosmos illuminated.[1]

We do not yet know if we can say the same for sentient life. In principle, purposive creatures could last as long as the heat remains on, until the last lumen has faded from the world's final supernova. But if the void of space is to be our guide, then the future of our universe may instead look much like its past: beautiful, but with no one to look upon it; vast, but with no one to explore; wondrous, with no one to wonder.

That is what this volume is about. Conscious experience is uniquely precious. In our universe, it is far scarcer than gold,[2] and it is valuable beyond any price. And remarkably, what we do today may determine the entire future of conscious life. Will we allow the light of awareness to fade into oblivion? Or will we fill the stars with matter that matters, for the eons that are to come? And should our descendents survive, will we bequeath upon them a bleak, grey crypt under the thumb of a bloodless tyrant? Or will we leave a jubilant world rich in art, in culture, and in splendour?

Such concerns lie at the core of the philosophy of longtermism, the idea that it is particularly important that we act now to safeguard future generations. How we act today will set the trajectory for the entire future, for all sentient beings who are yet to come. So how will we act, with the entire future at stake?

It is this imperative that inspires and animates The Long View. The volume's first three chapters articulate and defend the longtermist imperative at length. First is *'Expanding the Moral Circle to Future Generations,'* wherein Natalie Cargill lays out the history of humanity's *'moral circle,'* the imaginary boundary between those who are morally included and those who are morally excluded. Cargill argues that humanity's moral progress is not inevitable, but that it is possible for a small and dedicated group of

philanthropists to permanently transform humanity's moral circle for the better. Given this, philanthropists who want to make the greatest difference should work to expand the moral circle to include future generations.

Cargill's argument is developed and strengthened in *'The Case for Strong Longtermism,'* from Hilary Greaves, William MacAskill, and Elliot Thornley. These contributors argue that, if we want to do the most good we possibly can, we should take whatever actions we expect to have the most positive effects on the distant future, rather than taking those actions that are only beneficial in the short term. This is due to the vastness of the future: if future generations matter as much as present ones, then given how many future people may exist, anything we can do to put the future on a better long-term trajectory is vastly more important than anything we can do to improve the world today.

Lord Martin Rees, Astronomer Royal, offers a distinct justification for longtermism in *'Navigating the Next Century's Challenges'*. Lord Rees argues that the twenty-first century is the very first in which one species has the power to determine, for good or for ill, the future of the entire biosphere. We will do so as we navigate together several crucial challenges to the long-term future, including population growth, biodiversity, climate change, and advanced artificial intelligence. Humanity can navigate these challenges successfully, Rees argues, but only if we prioritise projects with a long-term political perspective.

The next four chapters take this thought as their point of departure. In *'Longtermist Institutional Reform,'* Tyler M. John and William MacAskill articulate a number of key reasons why political institutions around the world are so rarely able to adopt a long-term perspective, including uncertainty, bias, time inconsistency, and election incentives. They then propose and defend four institutional reforms that promise to substantially increase the time horizons of governments, so that they may rightly prioritise projects with long timelines: 1) government research institutions

and archivists; 2) posterity impact assessments; 3) futures assemblies; and 4) legislative houses for future generations.

Next are Lords John Bird and Robert Kerslake, who in their chapters outline the failures of long-term government planning in the UK and call for us to do better. In *'Lessons from the British Welfare State for Future Generations Legislation,'* Lord Bird hones in on the failure of the British welfare state, arguing that we could do much better in the future if we institutionalise future-oriented thinking in governments around the globe. It concludes with some reflections on the future we could achieve if we ensure that laws are fully and forensically assayed on their long-term future impacts, and an introduction to the Well-Being of Future Generations Bill put before the UK Parliament which may help to get us there.

In '*The challenge of effective long-term thinking in the UK Government and the critical role of philanthropy,'* Lord Kerslake compares the government's shortcomings with some of philanthropy's successes in adopting a long view. He tells the story of George Peabody, a philanthropist whose positive impact on the housing of the UK poor has outlived him for over a century, and argues that by supporting, working with, and learning from philanthropic actors, the government can achieve comparable successes.

Sam Hilton adopts a more optimistic angle on longtermist policy in the UK, telling the recent success story of the All-Party Parliamentary Group for Future Generations in his chapter, *'A Little Bit of Funding Goes A Long Way: The APPG for Future Generations'*. With only one year of a full-time staff equivalent, the APPG supported the creation of Lord Bird's Well-being of Future Generations Bill, launched an ongoing Inquiry on Longtermism in Policymaking, pushed the UK Parliament to set up a Select Committee on Risk Assessment and Risk Management and swelled membership to seventy-five UK parliamentarians. And the work of the APPG has only just begun to push for the fair inclusion of all future generations in government decision making.

The three chapters that follow explain cause areas of particular concern to policymakers and philanthropists wishing to protect future generations. One area concerns pandemics. The ongoing COVID-19 pandemic should prompt global leaders to take bold action to reshape international institutions and significantly invest to reduce future globally catastrophic biological risks. So Jaime Yassif argues in her chapter, *'Biosecurity, Longtermism and Global Catastrophic Biological Risks'*. To do so effectively, Yassif argues that we must maintain a broad perspective about the potential sources of such risks, addressing both naturally emerging pathogens and synthesised ones, and explains how we can do this.

A second area concerns extreme climate change. In *'Utilising Insurance for Climate Risk Reduction in the UK,'* Lord Des Browne charts a path forward for insurers to play a larger role in climate risk adaptation and mitigation. As Browne shows, insurers have an abundance of expertise, money, and perspective which they might bring to bear on this challenge, by serving as shock absorbers, risk engineers, and experts in risk modelling, pricing, prevention, and behavioral incentives. In this way the chapter identifies actions for government, the insurance sector, civil society, and philanthropists to take to improve collaboration and increase their collective influence.

A third area of concern to longtermists is the risks from advanced artificial intelligence. As Amanda Askell argues in *'Ensuring the Safety of Artificial Intelligence,'* artificial intelligence stands out among other forms of current technology as having particularly great promise and particularly great risk, and our work today to ensure the safety of artificial intelligence could have an enormous long-term impact on humanity. Such work includes gathering information about AI progress and impacts, investing resources into the safe development and responsible deployment of AI and working to resolve collective action problems that threaten to undermine these efforts.

The challenges facing future generations are enormous, and humanity is not currently well-suited to dealing with them. For this reason, Hiski Haukkala argues in the volume's final chapter that human society is in need of deep cultural change that stitches long-term thinking into the very fabric of our lives. In *'Traversing the garden of forking paths more wisely: The challenges of taking the long term into account in decision making,'* Haukkala surveys the obstacles to long-term thinking and proposes a new set of cultural traditions, which together will remind we who live today – along with all of our descendents – to be custodians of this planet and its long-term potential; to ensure a viable, open, and aspirational future for all generations.

Together, these contributions embody our heartfelt conviction that philanthropy need not relegate itself to an ameliorative practice of mere alms-giving. Every moment our entire future is at stake, and today's philanthropists uniquely possess the power, the opportunity and the moral obligation to protect posterity and maximise the flourishing of every generation yet to come.

# I

# The Longtermism Paradigm

CHAPTER ONE

# Expanding the Moral Circle to Future Generations

by
Natalie Cargill[1]
*Founder and CEO, Longview Philanthropy*

*Throughout most of recorded history, horrors such as chattel slavery, feudalism, and nationalist genocide were almost universally accepted. Today, we live in unusual times. Chattel slavery is abolished, half of all countries are democracies, and violent conflict is in decline. We have made a stunning amount of moral progress, in these domains and others. But, as this essay explores, this was not because the arc of the universe tends towards justice. On the contrary, our brains are hard-wired for prejudice and the universe is replete with examples of moral rise and fall, progress and collapse. Instead, a surprising amount of the advances we have made can be attributed to the work of a few moral pioneers – philanthropists and activists who took on audacious projects to not only help the needy, but to successfully challenge conventional wisdom on whose needs matter in the first place. In doing so, they expanded humanity's moral circle, benefited countless individuals who might otherwise have been neglected for centuries, and left today's moral pioneers with lessons in impact that we can learn from immensely.*

The first person to be described as a *'philanthropist'* in English[2] was the son of a Bedfordshire upholsterer, who failed to thrive as an apprentice grocer, and – like so many disaffected youths of his generation and ours – set out to sea in search of inspiration. Instead, he found pirates.

The Hanover had barely begun it's journey to Portugal when it was hijacked by state-sponsored militia from France. John Howard and his shipmates were stripped of their clothes, thrown into a dungeon, and left with nothing to eat for six days but a single joint of mutton, tossed onto the rancid floor for the prisoners to fight over.[3] Two further dungeons and untold torment later, Howard was released to the British in exchange

for a French officer. No sooner had he set foot in London, than Howard begged Commissioners of Sick and Wounded Seamen for help freeing his fellow captives that had been left behind. He described conditions in French prisons so barbaric that *'many hundred had perished'*,[4] and he was successful in securing their release.[5]

This was in 1755, when Howard was thirty. For nearly eighteen years afterward, his life returned to normal. Howard managed his estate, developed a reputation for unusual benevolence towards his tenants and the unemployed, and was eventually asked to serve as High Sheriff of Bedfordshire. His duties as a sheriff, however, included managing the local prison. What he saw on his first visit there left him shocked, horrified, and transported back to his nightmare as a French prisoner of war:[6]

> *Prisoners were chained by their neck, waist, hands and feet in heavy irons. Many were almost naked without shirts, shoes or stockings. They slept on the floor on dirty straw which was often so old it had turned to dust, unless they could afford to pay for a bed at 3s. 6d. a week (nearly two days' wages). If those on the floor wanted fresh straw, it cost them a penny a day. The walls were filthy and had not been whitewashed for years. Gaol fever killed prisoners regularly; so did hunger and cold. Boys as young as thirteen were confined along with hardened felons.*

Worst of all, jails throughout England were held privately, and instead of being financed by the state they stayed in business by charging the prisoners in their care for their time spent in jail. Given that prisoners had almost no way to make a wage, they were regularly held long after they had been acquitted by the state.

Appalled, Howard spent the next year visiting nearly every county jail in England and Wales. He scrupulously gathered evidence of the mistreatment at every prison and brought it before the House of Commons,

who were persuaded to pass two penal reform acts. The first required jails to be financed by the county, and the second required changes to 'improve conditions with measures for adequate clean water and sewage systems, better hygiene, and an upgraded diet.' Howard had copies of both acts printed at his own expense and sent to every prison in England.

But Howard didn't stop there. After the laws were passed by the parliament, he visited English jails a second time to see whether the laws were being enforced. They were not. So Howard decided to visit prisons elsewhere to see how English jails might be improved, traveling to Scotland, Ireland, France, Switzerland, Germany, Belgium, and the Netherlands.[7] He travelled relentlessly on horseback, averaging more than forty miles per day. Howard grew enormously disciplined in his project. Each day he rose at 3 a.m. to work. He disguised himself as a Parisian gentleman to evade arrest in Toulon, and as a doctor to gain entry to prisons. He became a vegetarian. By his own calculations, John Howard travelled a total of 42,033 miles in his pioneering efforts to reform European prisons.[8] He made over 350 visits to over 230 different jails and assembled an enormous amount of detailed evidence on these institutions, and published his findings in a 500-page volume entitled *The State of the Prisons in England and Wales*.

Howard befriended the utilitarian philosopher Jeremy Bentham, and together they made a great deal of noise about Howard's book. The book was widely read, in no small part due to Howard's financing of it from his own wealth. In total, Howard spent £30,000 (an inflation-adjusted £3.5m) of his own money on his travels and the distribution of his book.

John Howard's life ended in 1790, when he died of jail fever while traveling the world seeking a cure for the plague.[9] But his influence would outlive him for centuries. Howard's efforts led to his recognition as the pioneer of prison reform and a father of social science. His ideas took root in the minds of Sir Thomas Fowell Buxton and Joseph John Gurney,

the brother and brother-in-law of Elizabeth Fry, 'the angel of prisons.'[10] When Fry wanted to expose the prison system she followed Howard's example of compiling data in a book and bringing it to parliament, and her efforts eventually led to the passing of the Gaols Acts of 1823 and 1835.[11] To this day, criminals have extremely few legal protections, facing 'medical neglect, sexual abuse, lack of reproductive control, loss of parental rights, denial of legal rights and remedies, the devastating effects of isolation, and arbitrary discipline.'[12] But the pioneering work from John Howard and Elizabeth Fry paved the way for powerful criminal justice reform efforts continuing into the present day. Their successors in the UK abolished flogging, penal servitude, hard labour, and the death penalty, and passed prison sanitation, hygiene, diet, and health care reforms. Across the Atlantic, the ACLU's National Prison Project has won lawsuits on behalf of prisoners in more than twenty-five states, and 95 per cent of US citizens now support prison reform efforts.[13]

Conditions in prisons are slowly improving. But this was never inevitable. The vulgar conditions in European prisons had lasted for thousands of years, and were highly profitable, utterly accepted, invisible to the public eye, and subjected upon poverty-stricken and immigrant populations, who face great discrimination. Without Howard's pioneering efforts our own prisons today might yet look like those forced upon him by eighteenth century French pirates.

## 1. The Expanding Moral Circle

The work of the philanthropists who pioneered criminal justice reform dramatically improved the world for many millions of people – and perhaps for billions over the coming generations. They did so by extending the concern of reformers, of government, and of the public to the well-being of prisoners; by expanding humanity's '*moral circle*.'

We all have a moral circle: an imaginary boundary that we draw

around various objects in our world, or even around objects we might dream up.[14] Many things sit inside of that circle – your family and friends, probably all human beings, and likely all sentient animals – and many things fall outside of that circle – your car and furniture, your father's Roomba, and the insentient shambling dead of horror films. The objects inside your moral circle are the ones that you think deserve substantive moral consideration, in their own right, and which you would like to see flourish, for their own sake. Objects outside of your moral circle might still matter to you quite a bit – no one is eager to set fire to their savings account – but they mostly matter to you as instruments to benefit the things that are inside of your moral circle.

Today, nearly everyone would regard the conditions of eighteenth century prisoners as a moral atrocity. Thanks to the work of numerous advocates like John Howard, we regard all human beings as members of our moral circle rather than as mere instruments – even prisoners. But, as the example of British prisons shows us, things were not always this way.

We all know the idiom *'an eye for an eye and a tooth for a tooth.'* We're less familiar with the origins of that phrase. It's Babylonian penal code from the Code of Hammurabi, dated to 1754 BCE. But this is just the penalty for harming Babylonian citizens. As repugnant as the old adage is, the next line is far worse: *'If one destroys the eye of a man's slave or breaks a bone of a man's slave he shall pay one-half his price.'* This was compensation to be paid not to the slave, but to the slave's owner, to make up for their lost profit.

Although the Code of Hammurabi is the oldest known slave code, historically speaking it represents the norm rather than a deviation from it. Since the height of the Babylonian Empire, and probably much earlier, and until the eventual abolition of slavery in Western nations between 1794 and 1981, nearly every agrarian civilisation has accepted the practice of taking one group of people or another as chattel property.

All civilisations, at some time or another, have identified some group

of morally worthy individuals as undeserving of even basic moral consideration, treating them as mere instruments or things rather than as ends in themselves.

When we look at the world today, it is clear that we have a long way to go to secure the moral consideration that everyone deserves. It is incumbent on us to expand the moral circle of all human beings from its cruel and conceited beginnings to eventually include every sentient individual who lives today or who will live in the future. And yet, we live in a world where Black Americans are incarcerated at more than five times the rate of whites,[15] where five million people die from a lack of adequate medical care annually,[16] and where over one hundred billion animals are confined in factory farms at any given moment.[17]

But history gives us cause for optimism about the ability of dedicated philanthropists and activists to expand humanity's moral circle until all sentient individuals, present and future, are rightly considered. Our world has progressed from a period of universal chattel slavery to its total abolition; from widespread feudalism to nearly one hundred democracies.[18] And today, nearly half of women, globally, are paid for their labour.[19] The benefits of past efforts to expand the moral circle continue to cascade into the future, as we build on the work of our forebears to pursue a just world for marginalised communities. In this way, moral circle expansion is the ultimate movement-building project: today's movements for prison reform, and for racial, gender, and species justice simply would not exist were it not for the efforts of early pioneers like John Howard.

This is the key takeaway from Section 2: when we look at the history of human civilisation's first ten thousand years, we see that our moral circle is remarkably plastic, and can both expand and contract in different environmental conditions. The moral progress we've achieved was not inevitable, nor is our continued moral progress. But the past few centuries reveal that it is possible for a small and dedicated group of philanthropists to

transform humanity's moral circle for the better, as we will see in Section 3. The lessons from the previous sections are synthesised in Section 4: it is both possible and necessary that today's philanthropists act to expand the moral circle to future generations.

**2. The History of Humanity's Moral Circle (300,000 BCE to 1800 CE)**

It is commonly believed that the moral arc is long, but that it ineluctably bends towards justice. On this view, moral progress is simply inevitable, something that happens to humans rather than something that humans ourselves are doing, and could choose not to do. But this picture is wrong, for at least two reasons.

First, recent findings in developmental psychology show that human beings are hard-wired to make in-group, out-group distinctions, and to mistreat people who are different from them. In fact, our innate disposition to decide who is 'other' is one of the earliest cognitive faculties to develop in childhood. A study in PNAS last year showed a group of one-year-olds a vignette involving a victim being harmed by an aggressor, and a bystander coming to help. They found that the infants expected the bystander to help the victim when the two wore the same outfit, but the infants were shocked by the bystander's altruism when the bystander wore the same coloured outfit as the aggressor.[20] More unnervingly still, a 2013 study of 9-to-14-month-old infants watching a puppet show found that these infants preferred the puppets who shared their snack preferences to those who didn't, and they even preferred the puppets who punished other puppets who didn't share their same choice of snack.[21]

Evolution has innately primed us to like those who are similar to us, and even to wish harm on those who are different from us, beginning when we are mere infants. Without ongoing socialisation into a moral world-view where everyone is equal, we'll simply fall back into our native tendency towards prejudice.

Second, the history of humanity's moral circle does not reveal a simple picture of a moral circle ever expanding outward towards universal inclusion. Instead, it presents a messy picture of expansion and contraction, with no clear overall direction until the last three centuries.

Humanity's moral circle evolved as an early psychological trait to sustain cooperation in small communities and gain an advantage over competing human tribes. Early hominids survived by hunting together in groups and using their strength in numbers to ward off predators, and for this group altruism and group cooperation was essential.[22] Helping one's own tribe came at an evolutionary advantage: if the whole group cooperated together, provided aid when necessary, and punished uncooperative defectors, each cooperating member would reap the rewards of group cooperation, including safety, success hunting large mammals, and some resilience against the misfortune that might come from an unlucky week of hunting and gathering. But helping another tribe came at an evolutionary disadvantage, whether due to competition over resources and the control of the homo ecological niche, or simple opportunity cost. We don't know whether relationships between tribes were plagued with violence and resource scarcity or instead peaceable but distrustful,[23] but we do know that early human communities were closed and close-knit, and that moral concern was at first a scarce resource reserved in exclusivity for one's 150-500 closest friends.

For better and for worse, this small and unpromising beginnings of a moral circle proved to be remarkably plastic.[24] As ecological pressures and technological development transformed many hunter-gatherers into farmers during the Neolithic Revolution (c. 12,000 – 8,000 BCE), many humans began to develop large and highly-organised agrarian societies. Hunter-gatherers constantly on the move couldn't afford to have many children, but farmers could, so societies grew considerably. Agrarian societies required greater cooperation to store and disseminate surplus

goods when farming proved fickle, leading to more regimented social organisation and stratification.

Agricultural development led many human societies to expand, and with it, our moral circle: from small bands of up to five hundred to the entire local region. But it also led humanity's moral circle to contract in a number of ways.[25] The control and dissemination of surplus goods gave rise to class hierarchies with the creation of managers overseeing agricultural production. The availability of surplus goods led to the militarisation of human societies and an increase in violent conflict over property. And the benefits of upper-body strength for ploughing along with an increase in the number of children led to the origination of gendered labour and gender hierarchies, with males dominating the more highly-valued labour of farming and females relegated to childcare responsibilities.

The existence of warring city-states kindled stark moral tribalism between societies throughout the Bronze Age, the Iron Age, and Classical Antiquity. One of these warring city-states was Athens, where there remains a tombstone from the fifth century BCE that reads:[26]

> *This memorial is set over the body of a very good man. Pythion, from Megara, slew seven men and broke off seven spear points in their bodies ... This man, who saved three Athenian regiments ... having brought sorrow to no one among all men who dwell on earth, went down to the underworld felicitated in the eyes of all.*

Such radical exclusion of tribal outsiders from one's moral circle is representative of the attitudes of ancient human societies. According to Peter Singer, the Greek moral reformer Plato suggested an advance on this morality: 'he argued that Greeks should not, in war, enslave other Greeks, lay waste their lands or raze their houses; they should do these things only to non-Greeks.'

Around this time, and in the coming centuries, moral radicals such as Mozi, Jesus, Shantideva, Pythagoras, and Seneca advocated expanding the moral circle to every human being, or even to every sentient being. Many of these radicals had philanthropic sponsors – a group of three women funded Jesus's travels, and Shantideva was supported by monks – and they were doubtlessly influential: on reformers, scholars, religious communities; and eventually on their later followers, such as the Calvinist John Howard. But it was Plato who first got his wish, as tribalism gave way to nationalism, fuelled by imperial ambitions of mass colonisation. As agricultural societies grew and conquered neighbouring cities for resources, they expanded their ambitions of imperial conquest and began to form empires. Some of these empires enslaved their captives and formed feuds, while other empires allowed their colonies to self-govern, whilst extracting taxes and military might. The transition from localism to nationalism was an advance in human morality, but imperial expansion came at a dramatic human cost. In addition to more and larger-scale wars, at the zenith of imperial expansion came the colonisation of Africa, the extraction of 12.5 million slaves for forced labour,[27] and the advent of modern racism. The carnage of sixteenth century nationalism echoes into the present day. The economies of African countries colonised by Europeans continue to grow much more slowly than those of their neighbours, and the anti-Black racism initiated by imperial conquest persists through the centuries.[28]

### 3. The Civil Rights Era

Until the civil rights era the history of humanity's moral circle had been quite a disheartening mix of successes and failures. Far from a moral arc that bends gradually but inescapably towards justice, the first ten thousand years of human civilisation saw moral circle expansions – beyond the clan to the nation-state – but also enormous contractions – along lines of gender, race, and class.

But this all changed during the Civil Rights Era. The eighteenth and nineteenth centuries would see a dramatic change in the pace at which humanity's moral circle would enlarge. John Howard and Elizabeth Fry would successfully work to expand the moral circle to prisoners. Fearless Benjamin Lay would persuade the Quakers, and later William Wilburforce, to rail against slavery, using dramatic 'guerilla theatre' tactics – splattering fake blood on slave holders who came to Quaker meetings, and even briefly kidnapping the son of a Quaker family who kept a six-year-old girl as a slave, to show them what it felt like to have their child taken from them.[29] Mary Wollstonecraft would write *A Vindication of the Rights of Women*, decrying the status of marriage as a property relation, and calling for adequate women's education.

To this day we cannot be certain why philanthropists have been so much more successful at expanding humanity's moral circle during the last three centuries. Perhaps like dominoes, one successful social movement spurred on and energised another. Or possibly it was the gradual development of liberal individualism and the biological and cultural development of universalist morality that made it possible for these movements to take root. (This in turn might be explained by the medieval church's opposition to cousin marriages, which broke apart the family clan structure and paved the way for individualist societies.[30]) It might have been due in part to economic and technological changes, including increasing literacy rates to make movement-building possible. Or maybe it was all of these things.

Regardless of the reason, today's background conditions clearly make it possible for dedicated philanthropists to have an enormous positive impact by working to expand humanity's moral circle. For we have seen them do precisely this, throughout the 19th, 20th, and 21st centuries.

Arthur and Lewis Tappan and Gerrit Smith funded the American Anti-Slavery Society, along with many small donors mobilised by women

fundraisers, which would within two years become larger than the NRA or the Chamber of Commerce, relative to population, and turn abolitionism into a national crusade.[31] The Tappan brothers were nearly killed for their support of the society. Slavers placed a bounty of $50,000 on Lewis Tappan's life, if his head could successfully be delivered to New Orleans. But the threats to their lives only heightened their fervour, and the advocacy of the American Anti-Slavery Society stretched humanity's moral circle with a one-million tract civil information campaign that would be central to the movement to abolish slavery.

During the subsequent civil rights movement, Madame CJ Walker, a nineteenth century Black woman self-made millionaire philanthropist, would help seed the NAACP and fund numerous economic empowerment projects for Black women with the profits of her hair care empire. Andrew Carnegie and Julius Rosenwald would find inspiration in the work and life of Booker T. Washington, and contribute their fortunes to the cause of civil rights: with Carnegie funding Washington's Tuskegee Institute and Rosenwald providing fellowships to Black artists, scholars, and leaders such as Maya Angelou, W.E.B. DuBois, and Marian Anderson.[32]

In the 1960s and 70s, philanthropists would fuel numerous moral-circle-expanding legal victories around the world. The Albany Trust drove the 1967 passing of the Sexual Offenses Act, legalising consensual gay sex in the UK[33] In the US, the ACLU would win numerous anti-discrimination cases with their philanthropic funding. The Ford Foundation's major support for the ACLU's Women's Rights Project, led by Ruth Bader Ginsburg, would lead to victories in Califano v. Goldfarb, which struck down discriminatory elements of Social Security, and Craig v. Boren, a key litigative victory in the battle to establish gender discrimination as on a legal par with racial discrimination, and which entrenched a higher standard for evaluating sex discrimination claims that remains in place to this day. The Edna McConnell Clark Foundation became the majority

funder for the ACLU's National Prison Project, whose monumental victories we have already seen.

So while it can be tempting to believe the moral arc of the universe ineluctably bends towards justice, we are now in a place to see just how wrong this picture is. The moral arc of the universe is steamed and bowed by the work of philanthropists, litigants, and activists, who give selflessly of their time and their resources to find the most effective solutions to overcome marginalisation and oppression. Without the work of these moral pioneers, our world could be stuck permanently in an era of Athenian nationalism, putrid dungeons, and chattel slavery. These travesties persisted for thousands of years, and were highly profitable and widely accepted. It was far from inevitable, then, that we should have moved beyond them today.

**4. Expanding the Moral Circle to Future Generations**

The historical record shows the devastating effects that a contracted moral circle can have on the world. The past was truly terrible, marked by profound nationalism, militarism, and exclusion, and people literally owning each other as property. While humanity still has a long way to go, the present is a much nicer place to live, and this is due in very large part to the expansion of humanity's moral circle in recent centuries.

What lessons might we take away from our forebears, who fought courageously and effectively so that we might live in a more just and humane world?

First is the importance of taking on audacious, world-shaping projects with a long view of what we can accomplish. The work of John Howard, Benjamin Lay, and Mary Wollstonecraft took centuries to pay off in full, but they advocated on behalf of millions of their contemporaries and billions who would come after them, and a direct consequence of their work is that billions have and will continue to benefit from the immense

scale of their projects. This idea is echoed in the words of today's greatest philanthropists, such as Warren Buffett who writes that 'the only way you do significant things is by taking on tough things.'

Second, given the enormity of the projects that effective philanthropists take on, we cannot be put off by a need to take risky bets on huge successes. We don't know how many advocates before John Howard tried and failed to reform the world's prisons, but Howard's one success would have been worth the time of literally millions of advocates. Bill Gates writes similarly that *'philanthropy should be taking much bigger risks than business … If there are easy problems, business and government can come in and solve them.'* If philanthropists are to successfully shape the moral arc of the universe, doing what governments and markets are unable or unwilling to do, we need to take on daring projects to shape humanity's moral circle for generations to come.

Third, the greatest philanthropists think ahead of the moral curve, taking on neglected projects that no one else is willing to embrace. The philanthropists of the civil rights era railed against 'common sense,' fighting on behalf of individuals who no one else cared about. This meant that they were taking on the very largest problems, with the most need for reform, and it also meant that the problems they addressed may not have been solved by anyone else. Had they instead worked on crowded problem areas such as charity schools (*'in Fashion in the same manner as Hoop'd Petticoats, by Caprice,' said philosopher Bernard Mandeville in 1714, 'And the no more Reason can be given for the one then the other.'*[34]), then we might still be living with the neglected problems of stinking dungeons and chattel slavery today.

And finally, the examples of our forebears shows us the importance of expanding the moral circle. We have seen that the expansion of humanity's moral circle is not some supernatural phenomenon outside of the hands of individuals; humanity's moral circle has variously expanded and contracted

over the millennia, and our progress as a species has been a result of the advocacy of philanthropists and moral radicals striving for a better world.

We alone have the power to decide between alternate histories for our world, by choosing to accept moral exclusion within our societies or by advocating to expand humanity's moral circle to all who should rightfully be included. Given the opportunity before us to shape humanity's moral circle for the better, we therefore need to ask ourselves who is still being excluded, and what we can do to help them.

Criminals continue to have extremely few legal protections, despite the early pioneering work of John Howard, Elizabeth Fry, and the Edna McConnell Clark Foundation. In many US states, prisons are regarded as an exception to the thirteenth amendment to the constitution legally prohibiting slavery, and the treatment of prisoners falls so far below basic requirements of human rights that UN human rights documents such as the *Standard Minimum Rules for the Treatment of Prisoners* have been used successfully in legal trials on behalf of prisoners.[35]

Non-human animals, particularly those used in agriculture, have even fewer legal protections. Factory farming continues to be the dominant mode of food production, with over 90 per cent of farmed animals living and dying on Concentrated Animal Feeding Operations globally.[36] The industry standard typically involves confining these animals in cages so small that they are unable to stretch or turn around, and spend their lives in a state of stress and boredom while sitting in a pile of their own excrement. Factory farmed animals number in the trillions.

One further neglected group is of particular concern to this book. Future generations are utterly disenfranchised by governments. Only a handful of states have even informal representation for future generations, and only two nations in the world provide basic constitutional provisions for their security. Unlike many other disenfranchised groups, they are completely unable to participate in our markets, our movements, and

our civil society. In truth, they are entirely powerless today, and will live at the mercy of the choices of we who live in the present. Moreover, future people are vast in number. If we act wisely in the years ahead, our descendants could easily number into the trillions or beyond; given the long future that stretches before us, nearly everyone who will ever exist is likely to live in the future. And yet, as with the other groups considered in this chapter, future generations face exclusion from humanity's moral circle due to the mere circumstances of their lives and births.

Given the astronomical scale of the future, and the opportunity to make the future as marvellous compared to the present as the present is compared to the past, philanthropists who want to make the greatest difference with their efforts should work to expand the moral circle to future generations. To do so, we need to advocate for the representation of future generations within our legal systems (see John and MacAskill, Bird, Kerslake and Wagg, Hilton, this volume), the establishment of longtermism within culture (see Haukkala, this volume), and the growth of today's movement to protect future generations (see Rees, Yassif, Browne, Askell, this volume), and we need to promote research to inform all of our actions to positively shape the long-term future (see Greaves, MacAskill, and Thornley, this volume).

As with the moral visionaries before us – from John Howard to Madam C.J. Walker, and the philosophers of the ancient period – our efforts to expand humanity's moral circle to the marginalised will echo for centuries to come. If we protect our grandchildren, and ensure that our grandchildren protect their grandchildren in turn, we can ensure a flourishing future for generation after generation: a true sustainable investment to pay dividends from now until the end of recorded time.

CHAPTER TWO

# The Moral Case for Long-term Thinking

by
Hilary Greaves
*Professor of Philosophy and Director of the Global Priorities Institute, University of Oxford,*
William MacAskill
*Senior Research Fellow and Associate Professor of Philosophy, Global Priorities Institute, University of Oxford,*
and
Elliott Thornley
*DPhil Student in Philosophy and Parfit Scholar at the Global Priorities Institute, University of Oxford*

*This chapter makes the case for strong longtermism: the claim that, in many situations, our impact on the long-run future is the most important feature of our actions. Our case begins with the observation that an astronomical number of people could exist in the aeons to come. Even on conservative estimates, the expected future population is enormous. We then add a moral claim: all the consequences of our actions matter. In particular, the moral importance of what happens does not depend on when it happens. That pushes us toward strong longtermism. We then address a few potential concerns, the first of which is that it is impossible to have any sufficiently predictable influence on the course of the long-run future. We argue that this is not true. Some actions can reasonably be expected to improve humanity's long-term prospects. These include reducing the risk of human extinction, preventing climate change, guiding the development of artificial intelligence, and investing funds for later use. We end by arguing that these actions are more than just extremely effective ways to do good. Since the benefits of longtermist efforts are large and the personal costs are comparatively small, we are morally required to take up these efforts.*

The future is big. Our planet currently hosts around eight billion people. This century will see the birth of more than ten billion. If that number holds steady for just ten more centuries, we have a hundred billion people ahead of us. But humanity could last much longer than that. If all goes well, we can expect our descendants to outnumber us by

an even greater margin. We residents of the twenty-first century could turn out to be a drop in the ocean.

It is hard to grasp the size of humanity's potential. We are not used to thinking on the necessary time-scales. Harold Wilson said that a week was a long time in politics, and the remark seems true of many other domains too.[1] Our world changes so quickly that the consequences of our actions even a few years from now are tough to predict,[2] so it is no surprise that we rarely consider how our decisions might affect people living hundreds, thousands, or even millions of years in the future.

Nonetheless, we – the authors – believe that this neglect of the long-term future is a grave moral error. The view recently dubbed longtermism serves as a corrective.[3] According to this view, we should be particularly concerned with ensuring that the long-term future goes well. In this contribution, we argue for both longtermism and a further claim that we call strong longtermism which states that, in many situations, impact on the long-term future is the most important feature of our actions today.

Where exactly the short-term future ends and the long-term future begins is not important. We claim that the view is true even when we draw the line a surprisingly long time from now – say, a hundred years. The claim, then, is that the moral value of our actions depends primarily on their consequences arising more than a century in the future. That means that the predicted short-run value of our actions should not weigh heavily in our decision-making. Instead, our choices should be driven mainly by long-run considerations. Short-term effects matter, but they matter primarily as mediators of long-term effects.

We believe that strong longtermism has practical implications for individuals, charities, and governments.

Humanity's future could be extraordinarily valuable, and it currently hangs in the balance. If these facts were widely recognised, many of our priorities would change.

## The case for strong longtermism

The case for strong longtermism begins with the observation that our future could be vast. Astronomical numbers of people could exist in the aeons to come. Of course, the exact number is uncertain. The range of possibilities is wide. But, for our purposes, we can work with the expected number of future people. We calculate this figure in the same way that we calculate the expected value of a lottery ticket. Suppose that a ticket offers a 1 per cent chance of winning £300. Then its expected value is 0.01 × £300 = £3.

Whether a lottery ticket is worth buying depends on the numbers, and the same is true of our argument for longtermism. Here, as below, we will endeavour to be conservative in our estimates, erring on the side of underestimating the expected number of future people. If the case for longtermism is strong on these numbers, it will be even stronger on less cautious estimates. In that spirit, suppose that the chance that humanity survives until the Earth becomes uninhabitable – one billion years from now[4] – is just 0.1 per cent. The future is hard to predict, so being more than 99.9 per cent confident that we will not make it that far seems hubristic. Suppose also that ten billion people live in each century. In that case, the expected number of future people is at least 100 trillion (1014) – over 10,000 times the number of people alive today.

The size of that number might lead you to think that our estimates were not conservative after all. But note that the above calculation leaves out many opportunities for further inflation. Perhaps the most significant is the chance of space settlement. There are around 250 billion stars in the Milky Way, some of which will last for trillions of years.[5] If there is even a tiny chance that our descendants settle just a small fraction of these solar systems, the expected number of future people balloons upward.

Suffice it to say, the expected future population is large indeed. That is the first component of our argument for strong longtermism. The second component is a moral claim: all the consequences of our actions

matter. More specifically, the moral importance of what happens does not depend on when it happens. Agony and ecstasy occurring a hundred years from now matter just as much as agony and ecstasy occurring ten years from now.

This claim rules out what economists call a 'positive rate of pure time preference': preferring that good things occur at earlier rather than later times purely because they are earlier. To be sure, time preferences are appropriate in some domains. A pound now is preferable to a pound in ten years' time. But that is because we expect to be richer in the future, and pounds have diminishing marginal utility. Features of our lives that are intrinsically good or bad – things like joy and sadness – do not have diminishing marginal utility, so time preferences concerning these things are out of place. Consider an example. Suppose that you can save one person from torture ten years from now or two people from torture a hundred years from now, and that your decision will have no other consequences. It seems clear that, in this case, you should save the two people. Their pain should not be discounted simply because it occurs further in the future.[6]

Together, our argument's two components – the future is vast and all consequences matter – push us toward our conclusion: we can have a much bigger effect on the value of the future by trying to change its long-term rather than its short-term value. That in turn suggests that we should devote much more of our focus to considering the long-run effects of our decisions.

This claim is only strengthened by the observation that, so far, few people have recognised the importance of this longtermist insight. Most people and institutions are biased towards the short term.[7] If we direct our focus on the next few years, we enter a crowded field in which many of the best opportunities have already been taken and further progress is difficult. But if we instead cast our sights further, we find fresh ground.

Any opportunities here are less likely to have been taken, so we can expect to have an outsized impact.

**Longtermist initiatives**

But can we predictably improve the long-run future? One might think not, reasoning along the following lines:

Our world is so complex that it is impossible to foresee what effects our actions will have decades from now, let alone centuries. Since the long-run consequences of our actions are so uncertain, we cannot reasonably expect that any of our actions will make the long-term future better rather than worse. In light of this uncertainty, we should focus on the near future where effects are easier to predict.

We agree that the long-run value of many actions is hard to predict. But, importantly, this is not true of all actions. Some actions can be reasonably expected to improve humanity's long-term prospects, and this is enough to make strong longtermism true.

To explain one set of such actions, we first need to introduce an idea. Imagine a golf ball blown around a putting green by blustering winds. While the ball is on the turf, it will roll back and forth. The state of the scene will be constantly changing. But if the ball falls into a hole, it will remain there. The ball's being in the hole is what we call a persistent state. It is a state which, upon coming about, tends to persist for a long time.

Our world is like this windy putting green. It too has persistent states. Human extinction is one of them. The chances of humanity evolving all over again, post-extinction, are tiny. Human survival is another persistent state, albeit to a lesser extent. While the risks of extinction are real, there is at least a strong tendency for humanity to endure. These two persistent states differ in their long-run value. Our survival through the next thousand years and beyond is, plausibly, better than our extinction in the near future. So, if we can reduce the chance of human extinction,

we can predictably improve the long-term future.[8]

And it is increasingly recognised that we can reduce the chance of extinction. Matheny (2007), for instance, estimates that a $20 billion asteroid deflection system could halve the probability of an extinction-level asteroid hitting the Earth this century, reducing the risk from one-in-one-million to one-in-two-million. That decrease may seem small in absolute terms, but it makes an enormous difference to the expected number of future people. Recall that our calculation from the previous section gave us an expected future population of 100 trillion. Increasing the chance that this population gets to exist by just one-in-two-million is equivalent to saving 50 million lives in expectation. That comes out at $400 per life saved. And this is just one example. Combating other extinction threats – such as those arising from new or engineered pandemics – might be even more cost-effective.[9] This we could achieve by funding biosecurity work at the Johns Hopkins Centre for Health Security,[10] for instance, or the Future of Humanity Institute.[11]

That said, the case for reducing extinction risk hangs on our moral view. If we embrace a person-affecting approach to future generations – on which we care about making lives good but not about making good lives[12] – then extinction would not be so bad. It might even be judged good.[13] An asymmetric moral view – according to which bad lives get more weight than good lives – might lead us to a similar verdict.[14] And even on more standard views about the value of bringing new generations into existence, we might worry that future lives will be bad overall, so that extinction would be the lesser evil.[15]

Nevertheless, we argue, those drawn to person-affecting, asymmetric, and pessimistic views should still be strong longtermists. That is because extinction is not the only persistent state whose likelihood we can affect. Artificial intelligence presents another opportunity. Experts judge that there is a real chance that we develop advanced AI this century, with capa-

bilities exceeding our own across a wide range of domains.[16] As a result of their superior intelligence, these artificial agents may come to exert significant control over human affairs: making important decisions on behalf of individuals, governments, and other institutions. These agents might also endure indefinitely. Since their underlying code could be copied, they could outlast any given piece of hardware.[17] These two features of AI systems – their influence and their staying power – mean that they are likely to have substantial and lasting effects on the future.[18] That in turn suggests that we can have a beneficial influence on the long-term by increasing the chances that these systems are aligned with the right values. Work underway at OpenAI[19] and the Centre for Security and Emerging Technology[20] – to take just two examples – aims to achieve exactly that.

Another set of persistent states relates to climate change. A warmer climate could slow long-run economic growth, leaving future civilisation worse-off indefinitely.[21] It could also lead to the extinction of species, the destruction of coral reefs, and other forms of irreversible damage to our ecosystem.[22] Because these potential harms are near-permanent, we can expect that fighting climate change will have enduring effects on the future.

In sum, humanity finds itself in a delicate position. Our civilisation is currently poised between a range of persistent states. Falling into one of these states would likely have immense effects on the long-term future. Through the judicious use of time and resources, we can alter the chances that these states come about. As a consequence, we have the power to make our world better for generations to come.

But, as noted above, the case for strong longtermism hinges on the numbers. Not every lottery ticket is worth buying, and the same could be true of our proposed longtermist interventions. However, we argue that – even on conservative figures – the opportunities we list above are well worth the expense. Matheny's proposed asteroid deflection system

is one example. Another concerns artificial intelligence. If £1 billion of grants could reduce the chance of a catastrophic AI outcome – in which humanity's future is rendered near-worthless – by just 0.001 per cent, then a £10,000 donation can do as much good as saving 10,000 lives.[23] By contrast, it is widely agreed that the best available human-centric short-term interventions save roughly 4 lives per £10,000.[24]

And there is still much we do not know. Even if, at present, we cannot reasonably expect any of these longtermist initiatives to have better consequences than the most effective short-term actions, it remains possible that extra information would tip the scales in favour of a longtermist option. In that case, funding research into the long-run effects of various initiatives may be our best move. Since future people would likely take note of this research, we can expect our donations to increase the effectiveness of humanitarian efforts for many years to come.

Another option is to save our money.[25] We could set up a foundation or donor-advised fund with an explicitly longtermist mission. This fund would pay out if and when a good opportunity to shape the long-term future arises. The phase before the widespread deployment of advanced artificial intelligence would be one such opportunity. Since both the value of this longtermist fund and our knowledge about the efficacy of various actions is likely to grow over time, we can expect its impact to be especially substantial. The upshot is that we have a whole array of opportunities to benefit the generations who could exist in centuries to come. Even on the most cautious estimates, we should expect longtermist initiatives to do many times as much good as it is possible to do in the short term. So, if our aim is to do good, we should focus on the long term.

**Are we morally required to be longtermists?**

The argument above will motivate many people to set their sights on the long term. But others might want to hear more about the precise

moral status of longtermist initiatives. For even if longtermist actions like reducing extinction risk have the best effects on the future, that does not immediately imply that we are morally required to reduce extinction risk. For those people motivated mainly by a desire to avoid acting immorally, and not by a more wide-ranging desire to do good, this last step is important.

Let us expand on this point. On some moral views, doing what has the best effects is not always morally required.[26] Consider an example. Suppose that you are walking by a shallow pond, and a child asks you to wade in and get her football. You judge that the joy the child would feel in getting her ball back outweighs the frustration you would feel in ruining your clothes, so wading in would have the best consequences. Nevertheless, one might well claim, you are not morally required to wade in. You need not feel bad about staying dry. Similarly, one might argue, we are not morally required to do what has the best effects on the long-run future. We can instead devote our time and resources to other things.

Perhaps there are cases where we need not do what is impartially best.[27] We can allow that. But even so, we maintain that longtermist actions are morally required. This conclusion is implied by the following plausible claim: *When the action with the best effects has effects much better than other available actions, and any difference in personal costs is comparatively small, we are morally required to perform the action with the best effects*.[28]

This claim is compatible with the judgement that you need not wade into the pond. Although wading in would have better effects than staying dry, the effects are presumably not much better. The claim also makes sense of our judgements in an amended version of the case. Suppose instead that the child is drowning in the shallow pond. Then it seems undeniable that you are morally required to wade in and save her. Precisely because the stakes are high and the cost to you is small, you must do what is best.[29]

Our situation is closer to the drowning case than it is to the football case. Longtermist initiatives like preventing future pandemics are not merely slightly more effective than the best short-term initiatives. They are many times more effective.[30] Since the consequences of longtermist efforts are so much better in expectation, and the personal costs of a long-run focus are small, we are morally required to take up longtermist efforts.

**Conclusion**

Humanity's potential is vast and yet fragile. We could be on the verge of a long and magnificent future in which our descendants flourish for aeons to come. We could also be headed for an untimely end, or a drop into a permanent rut. Our fate is as yet undetermined. Influencing the chances that these futures come to pass is within our power.

These facts, in combination with a couple of plausible moral claims, have led us to a surprising conclusion: in many situations, effects on the long-run future are the most important feature of our actions today. This shift to a strong longtermist perspective is of no small importance. It directs us to spend significantly more of our time and resources on reducing extinction risk, preventing climate change, guiding AI development, improving institutional decision-making, fostering international cooperation, researching the long-run efficacy of various initiatives, investing funds for later use, and – almost certainly – many other things besides.

CHAPTER THREE

# Navigating the Next Century's Challenges

by
Martin Rees
*Astronomer Royal and Fellow of Trinity College, Cambridge*

*The twenty-first century is a crucial one. The Earth has existed for 45 million centuries but this is the first period in which one species, ours, is sufficiently dominant and empowered that it can determine, for good or ill, the future of its entire biosphere. This chapter provides an overview of the crucial challenges that humanity must navigate together to ensure that our long-term future lives up to its burgeoning potential, and introduces some promising solutions. First are challenges of population growth: by 2100, Africa's population could double twice, to 4 billion, creating a portent for disaffection and geopolitical instability. Second are threats to biodiversity – a crucial component of human wellbeing – from humanity's collective ecological impact. Third are problems of climate change and our necessary civilisational segue to a low-carbon future. Fourth are the promises and perils of AI. The lifestyle enabled by advanced AI seems benign, but may be imperiled by its threats to privacy and global economic equality. Humanity can yet navigate these crucial challenges successfully, but to do so it is imperative that we realise that we're all on this crowded world together and prioritise projects that are long-term in political perspective.*

First, the good news. For most people in most nations, there's never been a better time to be alive. The innovations driving economic advance – IT, biotech and nanotech – have boosted the developing as well as the developed world. Creativity in science and the arts is nourished by a wider range of influences – and is accessible to hugely more people worldwide – than in the past. We're becoming embedded in a cyberspace that will soon be able to link anyone, anywhere, to all the world's information and culture, and to most other people on the planet. Twenty-first-century technologies, in the coming decades, could offer

everyone a lifestyle that requires little compromise on what Europeans aspire to today, while being environmentally benign, involving lower demands on energy or resources.

This bright future is possible. But there is a gap between the way the world could be and the way it is actually heading – and that gap is growing. It's an ethical indictment of our world that its benefits are so unequally shared, that the wealth of its richest two thousand inhabitants could more than double the income of the 'bottom billion' – people struggling to live on only two dollars a day and with minimal access to health care, adequate diet, or education.[1] Equally, it's shameful that failures of governance lead to humanitarian disasters and conflicts.

The twenty-first century is a crucial one. The Earth has existed for 45 million centuries but this is the first period in which one species, ours, is sufficiently dominant and empowered that it can determine, for good or ill, the future of its entire biosphere. We've entered what some have called the 'Anthropocene' era. This chapter provides an overview of the crucial challenges that humanity must navigate together to ensure that our long-term future lives up to its burgeoning potential: challenges of population growth, mass extinctions, climate change, and advanced AI. To do this, it is imperative that we realise that we're all on this crowded world together and prioritise projects that are long-term in political perspective.

## A crowded world

Humanity's heavy collective footprint will get heavier still. Fifty years ago, world population was about 3.5 billion. It's now about 7.7 billion[2] The growth's been mainly in Asia; it's now fastest in Africa. The number of births per year is now going down in most countries.[3] Nonetheless, world population is forecast to rise to around 9 billion by 2050.[4] That's partly because most people in the developing world are young. They are yet to have children, and they will live longer. The age histogram in

the developing world will become more like it is in Europe. Moreover, despite welcome falls in infant mortality, the demographic transition towards a low fertility rate hasn't yet occurred in much of Africa.

The implications of population growth seem under-discussed. That's partly, perhaps, because of associations with eugenics in the 1920s and 1930s, and subsequently with Indira Gandhi's policies, and with China's drastic but effective one child policy. But also because doom-laden forecasts in the late 1960s – by, for instance, the Club of Rome – proved off the mark. As it's turned out, food production has kept pace with rising population; famines still occur, but they're due to conflict or maldistribution, not overall scarcity.

To feed 9 billion by mid century shouldn't be an insurmountable challenge. It will require further improvements in agriculture – low till, water conserving, and GM crops – and maybe dietary innovations, such as replacing energy-intensive beef with nutritious artificial meat.

To quote Mahatma Gandhi – 'enough for everyone's need but not for everyone's greed'.

Demographics beyond 2050 are uncertain – it's not even clear whether there'll be a global rise or fall.[5] Declining infant mortality, urbanisation and women's education trigger the transition towards lower birthrates – but there could be countervailing cultural influences. If, for whatever reason, families in Africa remain large, then that continent's population could double again by 2100, to 4 billion. Nigeria alone would by then have as big a population as Europe and North America combined.

Optimists say that each extra mouth brings two hands and a brain. But the geopolitical stresses are surely worrying. Those in poor countries now know, particularly via the internet, what they're missing – and they're not fatalistic about the injustice. Moreover, the advent of robots, and the 're-shoring' of manufacturing, mean that still-poor countries won't be able to grow their economies by offering cheap skilled labour, as the Asian

Tiger states did. It's a portent for disaffection and instability – multiple mega-versions of the tragic boat-people crossing the Mediterranean today.

Africa's predicament is aggravated by the loss of talent. Around half of its health workers want to leave[6] and their departure can be ill afforded; it's doubly tragic if, after moving to a developed country, they find they're not accredited, and doctors become cab drivers. It's just as bad in agricultural science, engineering, and all the other specialities that African countries require if they are to develop their potential. The poorest countries need to engage their diaspora communities, encouraging those with expertise to at least make regular visits back home. But wealthier nations should take some responsibility too. A cost-effective form of aid would be to establish, in Africa and elsewhere, centres of excellence with strong international links where ambitious scientists could work in less dispiriting conditions. They could then fulfil their potential without emigrating, and strengthen tertiary education in their home country.

There are encouraging initiatives. My distinguished former colleague Neil Turok spearheaded the African Institute of Mathematical Sciences in Cape Town – a pioneering institution that offers postgraduate courses to students from all over Africa (at far lower cost, incidentally, than could be achieved in Europe or the US). This is a model that has now been replicated in Senegal, Ghana and elsewhere. It would seem equitable, too, that for each skilled person 'drained' to the developed world, the receiving country should feed back sufficient resources to train two more.

Wealthy nations, especially those in Europe, should adopt policies like these. It's urgent to promote prosperity in Africa, and not just for altruistic reasons. Amidst the coming population boom, a thriving Africa is key to international cooperation and geopolitical stability.

**Mass extinctions?**

The risks of global population growth don't stop here: if humanity's

collective impact is too hard, the resultant ecological shock could cause mass extinctions.

Already, there's more biomass in chickens and turkeys than in all the world's wild birds. And the biomass in humans, cows and domestic animals is twenty times that in wild mammals. Overall, the biomass of wild land mammals is down 85 per cent in the last fifty thousand years due to the activity of humans, and plant biomass is down 50 per cent[7] Meanwhile, the rate of wild biomass depletion continues to accelerate.

Biodiversity is a crucial component of human wellbeing. We're clearly harmed if fish stocks dwindle to extinction. There are plants in the rainforest that might be useful to us. And insects are crucial for the food chain and fertilisation. But for many environmentalists, preserving the richness of our biosphere has value in its own right, over and above what it means to us humans. To quote E.O. Wilson: 'mass extinction is the sin that future generations will least forgive us for'.

Moreover, these pressures are aggravated by climate change – an issue that's not under-discussed, though still under-addressed.

**Energy and climate**

Despite the uncertainties – both in the science and in population and economic projections – there's a consensus that 'business as usual' scenarios, with continuing dependence on fossil fuels, could, by the end of the century, induce really catastrophic warming, and tipping points triggering long-term trends like the melting of Greenland's ice cap.[8] The updated IPPC report issued late in 2018 added urgency.[9] The pledges made at the 2015 Paris COP conference, with a commitment to renew and revise them every five years, are a positive step; the Glasgow follow-up conference in 2021 will be crucial, though the auguries are rather depressing.

Many still hope, nonetheless, that our civilisation can segue smoothly towards a low-carbon future. But the requisite policies are a hard sell for

politicians: they won't garner much enthusiasm for a bare bones approach that entails unwelcome lifestyle changes – especially if the benefits are far away and decades into the future. Indeed, it is easier to gain support for adaptation to climate change rather than mitigation because the benefits of the former accrue locally.[10] For instance, the government of Cuba, whose coastal areas are especially vulnerable to hurricanes and a rise in sea level, has formulated a carefully worked-out adaptation plan stretching for a century.[11]

Nonetheless, there is one measure to mitigate climate change that seems politically realistic – indeed, almost win-win. Nations should expand Research and Development (R&D) into all forms of low-carbon energy generation (renewables, fourth-generation nuclear, fusion, and the rest), and into other technologies where parallel progress is crucial – especially storage and smart grids. In all these areas, China could lead the world – deriving environmental benefit for itself, as well as economic benefit by dominating the international market for zero-carbon energy generators. And the UK, which contributes only about 1 per cent of global emissions, could do more for the world if it had a successful 'blitz' on relevant high tech – which should be prioritised as highly as medical or defence research.

The impediment to decarbonising the global economy is that, although renewable energy is getting cheaper, its intermittence means it has to be supplemented by some kind of storage, and that is still expensive. But the faster these 'clean' technologies advance, the sooner their prices will fall and become affordable to developing countries such as India, where more generating capacity will be needed, where the health of the poor is jeopardised by smoky stoves burning wood or dung, and where there would otherwise be pressure to build coal-fired power stations (burning cheap coal from Australia) .

If the Sun (or wind) is to become the primary source of our energy,

there must be some way to store energy, so there's still a supply at night and on cloudy days, or when the wind doesn't blow. There's already a big investment in improving batteries and scaling them up. Other energy storage possibilities include thermal storage, capacitors, compressed air, flywheels, molten salt, pumped hydro, and hydrogen.[12] The transition to electric cars has given an impetus to battery technology (the requirements for car batteries are more demanding than for those in households or large 'battery farms', especially in terms of weight).

We'll need high-voltage direct current grids to transmit efficiently over large distances. In the long run these grids should be transcontinental. They should transmit solar energy from North Africa and Spain to the less sunny northern Europe. More importantly, they should transmit energy east-west across the whole of Eurasia, to smooth peak demand over different time zones. This could surely be a globally-beneficial element of the Belt and Road initiative.

It would be hard to think of a more inspiring challenge for young engineers than devising clean energy systems for the world. Despite the ambivalence about widespread nuclear energy, it's worthwhile to boost R&D into a variety of '4th generation' concepts, which could prove to be more flexible in size, and safer than existing nuclear power plants. The industry, world-wide, has been relatively dormant for the last twenty years, and current designs date back to the 1960s or earlier. In particular, it is worth studying the economics of standardised small modular reactors which could be built in substantial numbers and are small enough to be assembled in a factory before being transported to a final location.

Attempts to harness nuclear fusion – the process that powers the Sun – have been pursued ever since the 1950s, but we're still very early in that story. Despite its cost, the potential payoff from fusion is so great that it is worth continuing to develop experiments and prototypes. The largest such effort is the International Thermonuclear Experimental

Reactor (ITER) in France. But there are now ten other smaller prototypes being built elsewhere in the world.

Why do governments respond with torpor to climate change, despite acknowledging it as the most prominent and pervasive environmental threat? Concerns about future generations (and about people in poorer parts of the world) tend to slip down the agenda. Indeed, the difficulty of impelling $CO_2$ reductions (by, for instance, a carbon tax) is that the impact of any action not only lies decades ahead but is also globally diffused.

The imperative for early action is stronger if we don't discriminate on grounds of date of birth, and value the life-chances of today's children, who can aspire to still be alive in 2100. Policymakers typically discount future benefits at a rate of 1-7 per cent per year. If we care about today's children, and the generations beyond them, we can't continue to discount future benefits (and – more importantly – future risks) at this rate. If there's no willingness to pay an insurance premium to safeguard against worst-case climatic scenarios, then the welfare of future generations is jeopardised.

**Computers and AI**

We owe the dramatic advances in computers to the fast-advancing ability to manufacture electronic components on the nanoscale, thereby allowing almost biological-level complexity to be packed into the processors that power smartphones, robots, and computer networks.

Thanks to these transformative advances, the internet and its ancillaries have created the most rapid penetration of new technology in history – and also the most fully global. Their spread in Africa and China proceeded faster than nearly all 'expert' predictions. Our lives have been enriched by consumer electronics and web-based services that are affordable to billions.

Machine learning, enabled by the ever-increasing number-crunching power of computers, is a potentially stupendous breakthrough. It allows

machines to gain expertise – not just in game-playing, but in recognising faces, translating between languages, managing networks, and so forth – without being programmed in detail. But the implications for human society are ambivalent. There is no longer a programmer who knows exactly how the machine reaches a decision. If there is a bug in the software of an AI system, it is currently not always possible to track it down; this is likely to create public concern if the system's decisions have potentially grave consequences for individuals.[13] If we are sentenced to a term in prison, recommended for surgery, or even given a poor credit rating, we would expect the reasons to be accessible to us – and contestable by us. If such decisions were entirely delegated to an algorithm, we would be entitled to feel uneasy, even if presented with compelling evidence that, on average, the machines make better decisions than the humans whose positions they have usurped.

Integration of these AI systems has an impact on everyday life – and will become more intrusive and pervasive. Records of all our movements, our interactions with others, our health, and our financial transactions, will be in the cloud, managed by a multinational quasi-monopoly. The data may be used for benign reasons (for instance, for medical research, or to warn us of incipient health risks), but its availability to internet companies is already shifting the balance of power from governments to the commercial world. Employers can now monitor individual workers far more intrusively than the most autocratic or control freak traditional bosses. There will be other privacy concerns. Are you happy if a random stranger sitting near you in a restaurant or on public transportation can, via facial recognition, identify you, and invade your privacy? Or if fake videos of you become so convincing that visual evidence can no longer be trusted?

The pattern of our lives – the way we access information and entertainment, and our social networks – has already changed to a degree that we would hardly have envisioned twenty years ago. Moreover, AI is

just at the 'baby stage' compared to what its proponents expect in coming decades. There will plainly be drastic shifts in the nature of work.

Clearly, machines will take over many jobs in manufacturing and retail distribution. They can replace many white-collar jobs: routine legal work (such as conveyancing), accountancy, computer coding, medical diagnostics, and even surgery. Many 'professionals' will find their hard-earned skills in less demand. In contrast, some skilled service-sector jobs – plumbing and gardening, for instance – require non-routine interactions with the external world and so will be among the hardest jobs to automate.

The digital revolution generates enormous wealth for an elite group of innovators and for global companies, but preserving a healthy society will require redistribution of that wealth. There is talk of using it to provide a universal income. The snags to implementing this are well known, and the societal disadvantages – such as diminished labour incentives – are intimidating. It would be far better to subsidise the types of jobs for which there is currently a large unmet demand and for which pay and status is unjustly low. It would surely be a win-win transition if those who worked in call centres or Amazon warehouses could find alternative employment in roles where human qualities like empathy were properly valued.

It's instructive to observe (sometimes with bemusement) the spending choices made by those who are not financially constrained. Rich people value personal service; they employ personal trainers, nannies, and butlers. When they're elderly, they employ human caregivers. The criterion for a progressive government should be to provide for everyone the kind of support preferred by the best off – the ones who now have the freest choice. To create a humane society, governments will need to vastly enhance the number and status of those who carry out care-giving roles; there are currently far too few, and even in wealthy countries caregivers

are poorly paid and insecure in their positions.

The lifestyle enabled by AI seems benign – indeed enticing – and could in principle promote Scandinavian-level satisfaction throughout Europe and North America. Moreover, citizens of these privileged nations are becoming far less isolated from the disadvantaged parts of the world. Because information technology (IT) and social media are now globally pervasive, rural farmers in Africa can access market information that prevents them from being ripped off by traders, and they can transfer funds electronically.

Recall, however, that these same technologies mean that those in regions of the world without developed industrialised economies are aware of what they are missing, decreasing geopolitical stability, and that they prevent still-poor African nations from undercutting Western labour costs to build capital. Developments in artificial intelligence and information technology will have numerous benefits for the already relatively well off, and indeed some for the global poor, but they nonetheless threaten to exacerbate global inequality.

**The long-term imperative**

How far ahead can we plan? There is a seeming paradox here.

For medieval Europeans, the entire cosmology – from creation to apocalypse – spanned only a few thousand years. And most of our Earth was *terra incognita*. But despite their constricted horizons in both space and time, they left a visionary legacy: masons built cathedrals, adding bricks to vast edifices that would take a century to finish, and which still inspire us almost a millennium later.

We've now mapped the Earth – and domains far beyond it. And we know that we're in a cosmos whose past and future are measured in billions of years. But despite our enhanced understanding of the natural world, and control over it, our planning horizon is shorter than it was

for our forebears: it rarely stretches beyond a decade or two.

At first sight it seems paradoxical that our hugely extended conceptual horizons – in both space and time – have been accompanied by shortened planning horizons. But there's a clear explanation. Despite living in turbulent and uncertain times, medieval people expected their children and grandchildren to live similar lives to their own. In contrast, we fully expect drastic changes in the backdrop to human lives, from one generation to the next. The unpredictability of the future is used as an excuse for downplaying long-term planning – even in contexts like population growth and global heating, where we can predict with at least some confidence beyond 2050.[14]

To meet the needs of the present – especially those of the poor – without compromising the ability of future generations to meet their own needs: these are the goals set by the 1987 Brundtland Report to the UN.[15] Environmental degradation, unchecked climate change, population growth, and unintended consequences of advanced technology could trigger societal catastrophe. But there's a collective failure to plan for the long term, and to plan globally.

We all surely want to sign up to reach this goal in the hope that by 2050 there will be a narrower gap between the lifestyle that privileged societies enjoy and that which is available to the rest of the world. This can't happen if developing countries mimic the path to industrialisation that Europe and North America followed. These countries need to leapfrog directly to a more efficient mode of life. The goal is not anti-technology. More technology will be needed, but channelled appropriately, so that it underpins the needed innovation. The more developed nations must get there too.

There are signs of hope. Polls show, unsurprisingly, that younger people, who expect to survive most of the century, are more engaged and anxious about long-term and global issues. Calls to effective philanthropy remind

us that urgent and meaningful improvements to people's lives can be achieved by well-targeted redeployment of existing resources towards developing or destitute nations. Wealthy foundations have real traction (the archetype being the Bill & Melinda Gates Foundation, which has had a massive impact, especially on children's health) – and even they cannot match the impact that national governments could have if there was adequate pressure from their citizens.

And we shouldn't downplay the role of the world's religions – often transnational communities that think long-term and care about the global community, especially the world's poor.

This is the first era in which humanity can affect our planet's entire habitat: the climate, the biosphere, and the supply of natural resources. Changes are happening on a time-scale of decades. This is far more rapid than the natural changes that occurred throughout the geological past; on the other hand, it is slow enough to give us, collectively or on a national basis, time to plan a response – to mitigate or adapt to a changing climate and modify lifestyles. Such adjustments are possible in principle – though there is a depressing gap between what is technically desirable and what actually occurs.

To respond effectively to the next century's challenges, and set humanity in good stead for the centuries that follow, governments need to prioritise projects that are longer-term than those born from the typical political perspective. That will require a widespread cultural change, but this, in turn, can and must be prompted by the efforts of a few. On that note, I give the final word to the great anthropologist Margaret Mead: 'Never doubt that a small group of thoughtful, committed, citizens can change the world. Indeed, it is the only thing that ever has'.

# II

# Policymaking for the Long Term

# CHAPTER FOUR

## Longtermist Institutional Reform

by
Tyler M. John[1]
*Head of Research, Longview Philanthropy and PhD Candidate in Philosophy, Rutgers University – New Brunswick*
and
William MacAskill
*Senior Research Fellow and Associate Professor of Philosophy, Global Priorities Institute, University of Oxford*

*In all probability, future generations will outnumber us by thousands or millions to one. In the aggregate, their interests therefore matter enormously, and anything we can do to steer the future of civilisation onto a better trajectory is of tremendous moral importance. This is the guiding thought that defines the philosophy of longtermism. Political science tells us that the practices of most governments are at stark odds with longtermism. But the problems of political short-termism are neither necessary nor inevitable. In principle, the state could serve as a powerful tool for positively shaping the long-term future. In this chapter, we make some suggestions about how to align government incentives with the interests of future generations.*

There is likely to be a vast number of people who will live in the centuries and millennia to come. Even if Homo sapiens survives merely as long as a typical species, we have hundreds of thousands of years ahead of us. And our potential is much greater still: it will be hundreds of millions of years until the Earth is sterilized by the expansion of the Sun, and many trillions of years before the last stars die out. In all probability, future generations will outnumber us by thousands or millions to one; of all the people who we might affect with our actions, the overwhelming majority are yet to come.

These people have the same moral value as us in the present. So in the aggregate, their interests matter enormously. Anything we can do

to steer the future of civilisation onto a better trajectory, making the world a better place for those generations who are still to come, is therefore of tremendous moral importance. This is the thought that defines the philosophy of longtermism.[2]

Political science tells us that the practices of most governments are at stark odds with longtermism. This may seem obvious. After all, governments are run by and for presently existing people; future generations have essentially no political representation, and even in the face of the catastrophic risk to future generations posed by climate change, governments the world over have failed to effectively respond. But the problems of political short-termism are even more substantial than they appear. Elected officials usually operate on 2-5 year time horizons, failing to look ahead even into the problems of the next decade. Estimates of the financial impacts of legislation typically extend to just a few years to a decade,[3] and politicians are rarely able to allocate time to agendas which do not bear fruit until after the next election. In addition to the ordinary causes of human short-termism, which are substantial, politics brings unique challenges of coordination, polarization, short-term institutional incentives, and more.

Despite this relatively grim picture offered by political science, the problems of political short-termism are neither necessary nor inevitable. In principle, the state could serve as a powerful tool for positively shaping the long-term future. Governments collectively spend over US$25 trillion per year,[4] and they are our best means of solving large-scale coordination problems. Moreover, research in legal theory and the social sciences shows us that countries' laws and policies have a profound effect on moral norms and attitudes.[5] The problem of aligning government incentives with the interests of future generations should therefore be a moral priority.

In this chapter, we make some suggestions about how we should best undertake this project. In Section 1, we explain the root causes of political short-termism. Then, in Section 2, we propose and defend four institutional

reforms that we think would be promising ways to increase the time horizons of governments: 1) government research institutions and archivists; 2) posterity impact assessments; 3) futures assemblies; and 4) legislative houses for future generations. Section 3 concludes with five additional reforms that are promising but require further research: to fully resolve the problem of political short-termism we must develop a comprehensive research programme on effective longtermist political institutions.

## 1. The Sources of Short-termism

The sources of political short-termism can usefully be divided into three major categories.[6] Epistemic determinants of short-termism are features of political actors' state of knowledge that prevent (even properly-motivated) actors from adopting appropriately longtermist policy. Motivational determinants of short-termism are features of political actors' goals and motivations that lead (even well-informed) actors to wrongfully discount the future. Institutional determinants of short-termism are features of political actors' institutional context that strip the political means from (even well-informed, properly-motivated) actors who could otherwise adopt more appropriately long-termist policy, or which make political actors less well-informed or less well-motivated. A lesson of this section will be that the causes of short-termism are myriad, and are ideally combated through a variety of reforms targeting different determinants.

The most widely cited epistemic determinants of short-termism involve rational discounting of future impacts because of a lack of information about the future.[7] When political actors are more uncertain about the possible benefits of an action due to uncertainty about causal mechanisms,[8] the future state of the world, the preferences of future people, or the security of political commitments,[9] then the expected value of those actions decreases relative to actions whose benefits materialise in the short term, which tend to be more certain.[10] Over longer timelines,

these problems proliferate, leading to greater discounting. While this discounting is rational, it could be reduced by increasing the availability of high-quality information about the future.

By contrast, irrational discounting primarily stems from cognitive biases and attentional asymmetries between the future and the nearby past. Cognitive biases include actors' tendencies to respond more strongly to vivid risks than to information acquired from abstract, general social scientific trends,[11] as well as excessive optimism about their ability to control and eliminate risks under situations of uncertainty.[12] The attention that political actors pay to the future and to the nearby past are asymmetric in that many simply forego the task of making predictions about the future and instead choose policies which have worked in the recent past. (As a topical example: when confronted with COVID-19 decision-makers may have assumed that the risks were similar to those posed by SARS and MERS, and based policy on that assumption, rather than making forecasts based on the properties of the novel coronavirus itself, such as its basic reproductive number and case fatality rate.) This is because predicting the future takes cognitive effort, but the past performance of policy is readily observable.[13] Voters have this bias too. It is easier for voters to base their decision on recent, observable track records than by gauging a candidate's preparedness for novel potentialities. Incumbent politicians know this, and therefore prioritise visible, short-term benefits which they can point to come their re-election campaign.

The literature on motivational determinants of short-termism has been dominated by discussion of political actors' apparent positive rate of pure time preference (our tendency to value near-term benefits more highly than distant benefits). While there is little consensus on the strength, shape, and malleability of political actors' time preferences, there is broad consensus that political actors have a positive rate of pure time preference and that this is a significant source of short-termism.[14] Political actors'

motivations are also made more short-termist by both self-interest and relational partiality. If political actors act to benefit themselves or their friends, family, or community, they will tend to privilege the interests of their contemporaries over future citizens, who are neither their friends, family, community, or themselves. Finally, numerous cognitive biases make political actors less motivated to care about the future, including problems of procrastination[15] and invisibility: our tendency to ignore problems that are not directly in front of us.[16]

Among institutional determinants of short-termism, election incentives are the most widely discussed.[17] Politicians strongly desire to be reelected – and parties desire to increase their immediate reputation – motivating them to prioritise policy which results in very near-term, visible benefits for which they can publicly take credit, while hiding or deferring costs. Politicians are also dependent on various firms and other bodies, whether for direct financial support or because they hold some other kind of influence. Where these bodies have short time-frames, and therefore short auditing durations, they exert pressure on political actors to use short auditing durations too.[18] And short auditing durations are institutionalised in numerous areas of policy. Performance indicators with short-term goals and positive discount rates, inadequate credit-tracking over longer time-frames,[19] and budget windows with short time-frames all incentivise political leaders to shift benefits to the short-term and costs into the future.

Beyond auditing incentives, there are also various pressures on careful deliberation. The 24-hour media cycle forces political actors to react to political issues almost instantly. Political polarisation significantly detracts from careful, collective deliberation due to the pressures to be uncooperative. Omnibus bills have further adverse effects on deliberation in that they are passed or rejected long before they can be carefully discussed in full. All of these pressures are particularly acute on issues with long time-scales, because there the situation is most epistemically

precarious, meaning there is more that can be contested as well as an even greater need for deliberate reflection.

The problem of time inconsistency also looms large among institutional determinants of short-termism.[20] A lack of strong commitment devices to ensure that governments will act on past promises leads to low levels of trust in long-term policy proposals.[21] When voters and influential elites cannot trust governments to act on their past commitments, they will oppose future-benefiting policy promises which might be reneged, as well as investment in future-benefiting policy which might be diverted to other ends. Finally, even when everything else goes well, institutions may simply be too weak to reliably bring about long-run outcomes or they may be plagued by collective action problems that undermine successful coordination.

## 2. Proposals

Responding to a variety of sources of short-termism across numerous areas and levels of government requires a variety of solutions. To illustrate the kinds of solutions we think would be viable responses to short-termism, and to advocate for these solutions in particular, we focus on four reforms: In-government research institutes and archivists, futures assemblies, posterity impact assessments, and legislative houses for the future. The first three are relatively moderate reforms that we think can be implemented right away, and which have strong evidential support. The last reform is much more tentative, but symbolises the kind of radical and highly under-researched reform we think longtermist political reformers should aspire to over the coming decades and centuries.

### *In-government Research Institutions and Archivists*[22]

Many sources of short-termism can be ameliorated through the production of digestible, widely-available, legitimate, and high-quality

information about future trends and the future effects of policy. We therefore propose that existing national governments invest in the creation of many new in-government research institutions with the express purpose of information-gathering and information-sharing about issues of long-term importance. They should be tasked with producing periodic, public reports that (1) chronicle long-term trends, (2) summarise extant research to improve its accessibility by the legislature, (3) analyse the expected impacts of policy, and (4) identify matters of long-term importance that fall outside of the political business cycle.

Various in-government futures research institutions have existed throughout the world, with varying degrees of success, including in the US and Singapore. Singapore's Centre for Strategic Futures has been influential in the civil service, and has improved the nation's receptivity to low probability, high impact events, such as global catastrophic risks.[23] The Office of Technology Assessment existed in the US from 1972-95, during which time it produced 750 studies on a broad range of issues from health science to space technology. A 1990 study by the Carnegie Commission on Science, Technology, and Government found that OTA reports were 'useful' to 'very useful' to 91 percent of congressional staff, and one analysis found that the OTA's 1980s studies on synthetic fuels 'helped secure approximately US$60 billion in savings.'[24] The OTA's elimination by the US Congress likely had a direct and harmful effect on the ability of congress to think constructively about future problems, and a number of policy writers and members of congress have advocated for reinstating it.[25]

Well-designed in-government futures research institutions can significantly reduce four major sources of short-termism. They can increase the appeal and robustness of long-term policy initiatives by decreasing collective ignorance about the future state of the world and about policy causation. They can reduce irrational discounting due to vividness effects and optimism bias by increasing the salience of possible future trajectories.

They can increase motivation to act for the long term among political leaders by bolstering liability mechanisms such as public disapproval and elections through the distribution of information to the general public. Finally, well-designed in-government research institutions are partially insulated from the institutional features that create a short-term 'political business cycle', allowing them to resist pressures to allocate agenda time only to short-term considerations.

The best in-government research institutions will generally be structurally and functionally independent of existing government offices, with the power to set their own research agenda, in order to insulate them from the political business cycle. It may also improve institutional independence to identify recruitment mechanisms which do not rely on the judgment of politicians, such as by tasking relevant professional associations with selecting researchers. The institutions must be given a very broad mandate – to report on all matters of long-term importance – both to ensure comprehensiveness and to give them the flexibility to adapt to changing circumstances over long periods of time. They should continuously engage with relevant academics and professionals from a range of backgrounds through incoming visits, paid consultancies, interviews, and events. Successful institutions might further be empowered to require reading and response from the legislature, ensuring that their advice is not ignored. Finally, in-government research institutions must work to improve the absorptive capacity of government, identifying and improving ways of summarising and packaging expertise so that it is readily usable for governmental decision-making.[26]

*Futures Assemblies*

To reduce the damaging influence of polarisation, short-term institutional incentives, and motivational failures, we propose the creation of a novel representative, deliberative, and future-oriented body: the futures as-

sembly. Futures assemblies are permanent citizens' assemblies with an explicit mandate to represent the interests of future generations. As citizens' assemblies, they are deliberative bodies of citizens who are randomly selected from the populace to provide non-binding advice to the national government on issues of long-term importance.

While no government has ever instituted a futures assembly similar to what we propose, citizens' assemblies have been employed for consultation and information-gathering purposes throughout the world. One of the most high-profile such initiatives was Ireland's 100-member Citizens' Assembly, which was established in 2016 and tasked with considering questions related to abortion, fixed term parliaments, referenda, population ageing, climate change, and gender equality. The deliberations of the Irish assemblies provoked a referendum to remove Ireland's constitutional ban on abortion and substantially shaped Ireland's Climate Action Plan.[27] The UK government's Select Committees have used citizens' assemblies on several occasions, most recently hosting a 110-member citizens' assembly designed to explore public views on strategies for reaching net zero emissions by 2050.[28]

Such real-world experiments, along with armchair evidence and a growing literature of 'laboratory' experiments suggest the promise of futures assemblies. Like in-government research institutes, futures assemblies would combat short-termism by providing permanent allocated agenda time to the consideration of the long-term future, providing a deliberative policy environment that is insulated from short-term institutional pressures. Because futures assemblies are explicitly tasked with the sole mandate of producing recommendations on behalf of future generations, we should expect that they will be much more long-term-focused than ordinary citizens.[29] While research institutes excel at producing high-quality information, citizens' assemblies excel in three other areas. First, because membership in a citizens' assembly does not

depend on election or successful fundraising, citizens' assemblies can almost completely eliminate short-term incentives from elections, party interests, and campaign financing. Second, citizens' assemblies have a demonstrated aptitude for reducing partisan polarisation and creating areas of consensus on matters of great uncertainty and controversy to enable timely government action.[30] Third, citizens' assemblies are statistically representative of the populace, positioning them uniquely to serve as a legitimate voice for the people. As a consequence, recommendations from futures assemblies will have an authority close to that of a consensus statement from the general population. Governments can ignore their recommendations only at a costly risk to their reputation.

The most promising futures assemblies would be relatively large (50-250 members) to ensure demographic representativeness and resistance to corruption from interests groups. To further aid against corruption and ensure representativeness and minimal resignations, assembly members should be paid a high salary, for example commensurate with the typical salary for members of the national legislative body. Assembly members should be empowered to call upon relevant experts, and to convene expert summits on matters of long-term importance. Full-time terms should be long enough for assembly members to build expertise but short enough to guard against disruptiveness and interest group capture, which we suggest is about two years. Ideally, assemblies would be empowered to set their own policy agenda, to further prevent capture by government interests, and their deliberations would achieve a very high level of publicity, to better enshrine their recommendations as legitimate and informally binding on the legislature.

*Posterity Impact Statements*

Posterity impact statements are another strong mechanism for creating political liability and gathering high-quality information about the long-

run effects of policy. These reports are functionally an extension of the environmental impact statements required by many governments for policy proposals with a potentially adverse impact on the environment. Our proposal is to require posterity impact statements on all proposed legislation with significant effects that occur beyond the ordinary two to four year policy window.

Environmental impact assessments (EIAs) are required throughout North America, Europe, and Australia. They are required of militaries,[31] developers,[32] state and local agencies,[33] and national governments.[34] Typically, EIAs are required when certain triggering conditions are met, such as when an action is likely to impact water, heritage sites, and other environmentally-zoned areas. As part of the EIA, the party assessed must identify and commit to a plan for reducing the adverse environmental impact of their actions. If the party fails to conduct an accurate EIA or to make good on their mitigation plan, they can be held legally liable for environmental damages.

While posterity impact assessments (PIAs) are a much newer idea, they are not entirely without precedent. The UK's 2020 Wellbeing of Future Generations Bill, started in the House of Lords by Lord John Bird, requires all public bodies to '(a) publish an assessment ("future generations impact assessment") of the likely impact of the proposal on its well-being objectives, or (b) publish a statement setting out its reasons for concluding that it does not need to carry out a future generations impact assessment' upon proposing any change in public expenditure or policy.[35] The impact statements must assess the impact of policy on 'all future generations… at least 25 years from the date' of publication, and include a statement of how any adverse impacts will be mitigated.

PIA requirements combat uncertainty about policy causation by requiring legislators to thoroughly research and publicise the long-term effects of their proposed policy for the opposing political party to scru-

tinise. They also hold legislators liable for the long-term effects of their decisions. Depending on the scheme, the associated liability mechanism can be "soft" in that it relies only on informal punitive and reward mechanisms, such as the embarrassment associated with putting forward a bill with harmful long-term effects, or it can be "hard" in that it is backed by formal sanctions, such as the requirement that legislators pay an insurance premium to cover expected damages. Both hard and soft liability mechanisms impose costs on legislators putting forward bills with adverse long-term effects, and so incentivise policymakers to be proactive about mitigating long-term harms. Ideally, they would also reward legislators putting forward bills with beneficial long-term effects, since these benefits may otherwise be unknown to legislative proponents or covered up by detractors. One simple such mechanism would allow expected benefits to offset a bill's expected future costs.

Posterity impact statement requirements should have triggering conditions and enforcement mechanisms which ensure that they are required in any conditions where posterity is affected, positively or negatively. The bill in front of the House of Lords ensures that PIAs are triggered on appropriate occasions by making them universally required, but there are various other triggering conditions that may suffice: PIAs could be required on submajority vote of the legislature, or upon order of a court. Ideally, PIA policy should require a zero rate of pure time preference and an open-ended assessment period. Significant impacts on future generations should not be treated as null merely because they are centuries away; we should ignore these effects only when there is no reason to think they are more likely on the proposed policy than its alternative.

*Legislative Houses for Future Generations*

The three reforms just proposed have been relatively moderate, soft-power political reforms with payoffs that are potentially quite large. The

reason for this is straightforward: moderate, soft-power reforms can feasibly be implemented immediately and have a lower likelihood of being repealed when the government changes hands. The recent examples of Hungary's (2008-2012) and Israel's (2001-2006) Commissioners for Future Generations suggest that more powerful institutions that hold veto or other similarly decisive powers are currently too partisan to survive an election cycle.[36] To pave the way for the powerful future-oriented institutions that longtermism recommends, we may first need to engage in more modest reform efforts to signal the importance of the long term and lay the groundwork for more vigorous possibilities.

Over the coming decades and centuries, however, longtermists should consider much stronger institutional reforms that can transform governments into the kinds of institutions that can positively shape the future on very long timescales. While it is currently difficult to imagine exactly what sorts of institutions could do this, we propose one possibility: an upper house in the legislative branch of government devoted exclusively to the wellbeing of future generations.

In the system we envision, bicameral national legislatures would be constituted by a lower house focused on attending to the interests of the people who exist today and an upper house focused on attending to the interests of all future generations. Legislation may be proposed in either house, but must be passed by both houses to become law. Thus, each house would provide a check on the other, ensuring that neither future-oriented nor present-focused legislation can be dominant. A strong constitution providing basic rights and freedoms to both presently-existing and future people would provide another strong backstop against the tyranny of either house.

Two major questions are relevant to the design of a successful legislative house for future generations: who serves?, and how do we ensure they have the right incentives? While we cannot provide conclusive answers

to these questions, we have some preliminary ideas about what design might work well. Random selection of legislators from among voting-eligible citizens may provide the best mechanism for deciding who serves, given its aforementioned elimination of short-term incentives from elections, party interests, and campaign financing, as well as its ameliorative effects on industry corruption and partisan polarisation. A subset of the legislators might be selected at random from among eligible experts, stratified by area of expertise, in order to ensure technocratic competence across a range of issues.

To ensure that the House has the right incentives, we suggest three further mechanisms. First, the House should have objective and concrete long-term performance metrics which are set in close deliberation between the House and an informed and non-partisan body, such as an independent research institution for future generations. These metrics should be updated regularly to correct for prediction errors and new developments. Second, the sole constitutional mandate of the House should be to set and pursue the achievement of long-term performance metrics. This would have some effect on the way House legislators conceive of their work and on the kinds of public justifications they can offer for their actions: any justification given to the media or in proposed legislation must cite concrete performance metrics. Third, the House should employ backwards pensioning: the pensions of House legislators should be determined some specified number of decades in the future, based on the House's long-term impacts. One obvious way of evaluating the House's impacts is by the extent to which objective performance metrics have been satisfied in the decades after their rule. An alternative evaluation mechanism would adjust pensions based on the retrospective attitudes of the future generations house in power at that future time. In this case, the reward scheme could have an intergenerational chaining effect. In deciding the pensions of past legislators, each house would be incentivized to consider

how their pension choice will be evaluated by those who will in turn reward them, decades into the future, thus providing incentives for every house to consider the long-term impacts of their decisions. Regardless of how pensions are decided retrospectively, the mechanism suggests an age limit on selected legislators to make it probable that they each live long enough to collect and enjoy their adjusted pensions.

This proposed reform is speculative, and to work effectively it would require both robust future-oriented research institutions and a long-term-orientated culture stronger than we find in any modern nation. Nonetheless, we hope that it symbolizes the kinds of powerful and imaginative political reforms that we should aspire to in the years ahead, and serves as fodder for much-needed additional research on longtermist institutional reform.

## 4. Future Directions

We have proposed several longtermist institutional reforms that can be implemented in the near-term future – in-government research institutions and archivists, futures assemblies, and posterity impact statements – and we have gestured at the more radical (but we think entirely warranted) reform of even having a separate, future-oriented division of government.

While the reforms proposed are significant, and will help to put society on a better long-term trajectory, we see this discussion as being merely a first step on a long path toward truly longtermist political institutions. The movement for longtermist political reform will require substantial advocacy. It will also require much more research. Other promising possibilities which require further research include longer election cycles to reduce perverse election incentives,[37] novel commitment mechanisms to enable longer-term decision-making, extra votes for parents to use on behalf of their children (or "Demeny voting"),[38] taxation for long-run negative and positive externalities, and broader long-term pay-for-

performance incentive schemes such as tying public pensions to national performance. Because the literature on political short-termism is young and still relatively conservative, there are likely to be many more promising possibilities that we have not yet uncovered.

The indeterminacy of the future and the complexity of policy systems can cause a sense of vertigo when considering the possibility of longtermist institutional reform. But the sorts of societal change that the more enlightened of our forebears envisaged – the suffrage of women and people of colour, or the protection of the natural environment – must have seemed no less giddying. Even if future generations can never truly participate in our political system, through progressive changes to our political institutions we may one day give them the consideration they deserve.

CHAPTER FIVE

# Lessons from the British Welfare State for Future Generations Legislation

by
Lord John Bird MBE
*Co-Founder, The Big Issue*

*To best ensure the wellbeing of future generations we must look to the successes and failures of past future generational thinking. The greatest expression of this is the creation of the British welfare state. The welfare state was created to address the iniquities of unemployment, poor housing, poor education and poor nutrition for future generations. It was a poverty alleviation programme of the deepest kind. Unfortunately, because the welfare state was so limited in its ability to create the change that was sincerely desired, over 70 years later we are now living with the costs of its failures. This chapter outlines the failures of the British welfare state and argues that we could do much better in the future if we follow the Welsh example of institutionalizing future-oriented thinking in governments across the globe. It concludes with some reflections on the future we could achieve if we ensure that laws are fully and forensically assayed on their long-term future impacts, along with an introduction to the Well-Being of Future Generations Bill before Parliament which may be able to help get us there.*

If we want to ensure the wellbeing of future generations it is essential to look at former future generational thinking. A great example of this is the creation of the British welfare state. One reason we need to improve future generational thinking in government is because we have not got the future right in the past. We extrapolated poorly understood trends of the time into the future rather than thinking clearly about what would actually be needed.

I was born two years before The Future began. I was a part of a future generation. The future then, launched in 1948, was called The Welfare State. It still exists in parts, the NHS probably the most famous part of it,

still clinging on with its 'Health for all' mantra. The educational dimension clings on also. But neither the NHS nor state education has brought the health or the education for all that was promised over seventy years ago.

Being a baby of the future in 1948 when the welfare state was put together with a rather ad hoc, well intentioned mixture of middle class prejudice, sentimental attachment to the poor, and shoddy intellectual tools, was to face a future that was always going to be lumpy.

The welfare state was created to address the iniquities of unemployment, poor housing, poor education and poor diet. It was a poverty alleviation programme of the deepest kind; something that had never been attempted in the UK before. Because that social creation was so limited in its ability to create what it sincerely desired, over seventy years later we are now living with the costs of its failures.

We have a social security system that does not provide security; a health service so full of people who have passed through poverty that it is almost full – a vast sponge absorbing those whose condition is the result of generations of poverty thinking and poverty living.

One of the first things the architects of the welfare state failed to address was the enormous class division that existed in British society at the time. Britain was an incredibly poor country for most of its occupants. You lived hand to mouth, in poor housing, with poor hygiene and poor food. Over half the living accommodation of the working classes was substandard. A large part of the working population was illiterate, lacking the education and skills of the middle and upper classes.[1]

It was a world that was demanding reform. The vast majority of people worked in manual jobs that did not lead to an improved lifestyle through promotion and skill enhancement. Workplace accident and injury was rife. Diseases related to the extraction and manufacturing industries – coal, steel, construction – were normal. You went to a form of work that made you ill.

The welfare state was set up specifically to aim for a new future without the above list of human failures. But it did not address the deeply entrenched nature of poverty and the deep psychological problems living in poverty throws up. Possibly one of the most criminal missed opportunities in creating the welfare state was in education. The reform of education led earlier to extending the age at which a student could leave school from 14 to 15. The creation of secondary modern education for the majority was the result. But the creaming off of about 11 per cent of high-achieving working class students into grammar schools laid the basis of social divisions that we are still suffering from.[2]

The future was largely planned to have an unskilled or semi-skilled working class that carried out manual labour until they were too worn out to continue. Today, with about 70 per cent of government and parliamentary time, as well as that of local authorities, given over to dealing with the collateral damage thrown up by poverty, you can spot a result of poor future planning.

If you were to question the vast number of people in prison, the working poor, those on long-term unemployment through health or a lack of jobs, those that fill up our A&E departments, you would find out that they did poorly at school. Probe further and you will find that their parents, and their parents' parents, likewise did not excel at education.[3]

Our existing poor are largely the grandchildren left behind by the future generations planning of the welfare state. They are those who were not skilled-up at school or through training and higher or further education; those who were only offered manual, health-eroding jobs; those who were not given the chance of self-improvement, the self-improvement that was offered to those who passed their eleven-plus exam and moved into grammar school.

The law of unintended consequences abounds. It has resulted in a largely under-educated, under-skilled workforce moving towards a world

where manual labour and semi-skilled jobs are being whittled away, and where jobs demand more skill rather than less.

You can't entirely blame the welfare state for its failings. What an enormous task it would have been to completely upskill the majority of working class kids. And bear in mind that the culture of education would also have had to be instilled into home life for it to take root. You would have had to break the previous generation's own prejudices against education that existed in many working class homes.

There would have had to be an enormous reeducation among parents as much as among children, teachers and school heads. The kind of energy and resources needed would have had to be massive. Hence the desire to have a transition from manual labouring into new kinds of jobs was never really fulfilled. Britain really did win the war and lose the peace, in the fullest sense.

**The UK on its back**

The UK after the war was on its back. It owed the US treasury billions of dollars. The debt accumulated through fighting the Second World War was not finally paid off until 2005. With poor housing, poor health, poor education and poor jobs, how would it be possible to get out of this trap? The country's roads, rail and distribution networks were antiquated and ill-fitting the requirement of modern industry. Its basic nationalised industries had suffered through under-investment by their former owners and were full of Victorian working practices and machinery.

How could it get out of this trap when the government did not invest in new industries but kept the old industries of extraction and manufacturing going on in the same way? None of the basic industries would have survived without government subsidies that were introduced in the First World War and persisted until the mid 1980's when Margaret Thatcher brought them to an end.

The US taxpayer, frightened of Soviet expansion into Europe and the far East, invested in recovery. France, Germany, Italy, Japan, and others were given money through The Marshall Plan to get their economies back working. Germany, Italy, and Japan spent the money they received on creating new industries and new jobs. Germany in particular upskilled its workers and concentrated on improving education. The UK was also given Marshall Plan support, but they spent it on maintaining their empire and creating the welfare state.

As part of a former future generation I am determined that our work on behalf of future generations does not make these same mistakes.

The COVID-19 pandemic has thrown up many problems that reflect on the inheritance left by the welfare state experiment in the postwar years. The fragility of the NHS is one. Because the NHS has had to pick up the social as well as the health bill of poverty and an aging population, it has always operated at near full capacity.

Many of the problems that it must deal with today are the problems of a class system that was never removed by the welfare state. This class system is a division between the well fed and the nutritionally deficient. It is between the well remunerated and the underpaid, working poor. Between the adequately housed and the poorly housed. Between the highly educated and the under educated. It is between a world full of wellbeing and a world where wellbeing is a luxury.

By looking closely into the entrails of former legislation for future generations we might well learn to do better. Can we do better?

In 2015 Wales passed their Well-being of Future Generations Act. The act places a duty on all public bodies in Wales to create and carry out long-term wellbeing objectives, and establishes a Future Generations Commissioner to ensure compliance with the mandate.

I look at this act with interest. Could it stop the persistent and predictable failure of the poor due to a lack of education and social opportunity?

Could it address the uglinesses thrown up by the social divides between the comfortable at the struggling discomforted?

Could the Welsh example help with the rejuvenation of cultural and political thinking that embraces social justice, and not mere wishful thinking and hope?

If we had a Future Generation commission in the UK in 1948 would it have been able to see the dangers of creating a low wage economy, based on manual jobs that were likely to disappear in a generation? Of an under-skilling of working class people, starting at their schools and ending in their typically dead-end jobs that were all that was on offer once they were old enough to join the workforce?

Would it have been able to suggest an alternative course to creating a low wage, low investment economy, that ends up with a service economy that will now see hundreds of thousands out of work, post pandemic?

The Welsh example and the history of the welfare state show us that a completely new mindset needs to be embedded in the political process.

They show that short-term thinking cannot be allowed to hold such political sway; that decisions in the present must anticipate the social echo they cast forward. I am convinced that if we can introduce acts to the rest of the world based on the kind of thinking behind the Welsh example, we can finally make securing the wellbeing of future generations a reality.

One of the things that first drew me to the Welsh Future Generations act was its fit with the kind of work The Big Issue had been groping towards. We had become increasingly aware that however many people we helped out of homelessness there were always more to take their place. Homelessness was often created almost at birth; a kind of perverse birthright. When you look at the ingredients that make up most peoples' homelessness there are normally failures in early life – around family, education, employment and a lack of wellbeing.[4]

We became more obsessed with trying to break the constant flow of

people, failed by the welfare state, who were now coming to us. Prevention became an increasing interest. What would you need to do to prevent people becoming homeless? What would you have to improve in society to achieve that end?

It all came back to government. It all came back to legislation. We noted that when The Big Issue started in the early 1990s there was a very large amount of very young people sleeping rough in London. Thousands who had left their homes and fallen homeless, reduced to begging and sometimes stealing or selling themselves as sex workers. We wanted to know how they had come to sleep in doorways.

The answer was changes in social security provision made in the late 1980s. In order to get young people moving from the home they shared with parents who were on social security, the state stopped their benefits. Up until then sixteen to eighteen year-olds were given their own benefit, even if they lived at home with their unemployed or sick parents.

In one fell swoop they stopped the benefits, and stopped children making a contribution to the family purse. Hence many left or were thrown out. And thousands of them slipped into homelessness. This was the corollary of trying to get young people 'moving', not understanding that it would lead to a vast increase of young people becoming homeless.

What became apparent to me was the need to actually get into parliament and try to influence legislation. Too many times I was being described as someone who knew how to 'think outside the box'.

Of course, you need people thinking outside the box if the box doesn't work. But the box is where laws and acts are created. It was there that change could be made to curb the unintended consequences destroying part of a generation of young people and get things moving in the right direction. So I applied to join the House of Lords as a prevention-leaning crossbencher, allied to no party. I was determined to stop injurious legislation that often harmed those who most needed our help.

Becoming aware of the future generations work of the Welsh government was a great source of joy. Here Wales was leading the way, showing that government was capable of real long-term vision.

**The future we could have**

What do I want for future generations? What do I want life to be like for all? Certainly, I want to end the kind of housing that I was born into. Slums, the kind I lived in, have gone, only then for poorly made housing to become the new slums. Thousands upon thousands of slums were pulled down in the 1950's only to be replaced by social housing that was substandard and poorly thought through. Those also have been pulled down. High-rise hells were put up that destroyed community life. We have to ensure we do not go down the road of putting up cheap housing for short-term use.

For future generations, I want to see the end of the 35 per cent of children leaving school having failed in their education. There to fill up the working poor jobs of today.

I want us to make sure that children are not left outside of social provision, so that when they get to school they are bound to fail.

Future generations cannot thrive unless we address the most pressing environmental issues along with the social. Among other things, good, freely available education will ensure that nature is understood and treated better by all. We have to bring nature and the environment into the lives of the many. At the moment it does not appeal to many because poverty closes down relationships with matters of little direct relevance to one's immediate survival.

If we embrace environmental issues and make them central to our lives this becomes a device for social and educational enhancement. Nature is a vital means of lifting people out of the throwaway, cheap consumer tastes of today.

But most crucial is to bring a mindset change, a mindset change that the Welsh future generations legislation has found hard to instigate. Changing minds about education and social justice, and about the need to think long-term, are the biggest obstacles. Politicians and legislators have to understand that no law should be accepted that poses risks to the wellbeing of posterity. That is an uphill struggle.

It is a struggle that we cannot avoid. It was the same kind of struggle that was necessary for my parents' generation, who could not necessarily see the advantages of education, and who believed that the lives of their children would simply be a reproduction of their own lives.

We have to work hard to ensure that laws are fully and forensically assayed on their future impacts, and at the same time concentrate on changing the minds of those who must deliver; in government or in local authority.

I certainly do not have a blueprint for the future that is more than a sketchy affair. But I do believe that if the importance of long-term thinking can be enshrined in law, then we can finally begin a full appraisal of long-term effects to bring about a different future for the coming generations.

As the case of Britain's welfare state shows us, the path to long-term thinking is not easy. There are many pitfalls. We must accept the fact that the road to hell is paved with good intentions. That the founding mothers and fathers of our welfare state started from the best of intentions; to stop tomorrow being a poor repeat of yesterday and today.

That is why I have proposed the Well-Being of Future Generations Bill, a bill inspired by the Welsh example and for which we must rally to become an act. Among other things, the Well-Being of Future Generations Bill follows the Welsh Act in placing duties on public bodies to set and work towards long-term wellbeing goals, and in establishing a Future Generations Commissioner for the UK to oversee UK policy and duties. It requires ministers to publish risk assessments and trend reports,

and establishes a Joint Parliamentary Committee for the Future to hold government ministers accountable for short-termist decision-making.

For if today is not passed through the prism of tomorrow then a debris of social dislocation for future generations to wade around in will be our only legacy. Concerns for Future Generations are seen by some as a luxury. As if we can't afford to factor the future into today.

Imagine what it would have been like if we had the sagacity back in 1948 to not burden the future with so many failed lives. Where poverty ruins the taste of living for too many. Where the just-about-managing and the working poor live their lives in the shadow of democracy and never in its full light.

The Welsh Future Generations Act is one of the most important pieces of legislation enacted in modern times. Its commitment to the new-born and the unborn, to future generations, makes it a standard for governments around the world.

Such a commitment cannot be dodged. With our current world turning into a very large tub of toxic waste, our seas and our air poisoned with pollutants, it is clear what damage poor concepts of the future can do.

Of course, we need to redouble our efforts to clean up the place that we live in. But we also need to rally for the future to ensure our legislators don't destroy tomorrow for the short-term gains of today.

I entered parliament to get rid of poverty. Poverty is the waste product of poor politics and poor laws. It is the waste product of a society that has not imagined the damage that can be done by doing the wrong thing to tomorrow. I am sure you are as sick as me of wading around in the debris of past poor thinking. That is why we have to embrace future generations today.

CHAPTER SIX

# The Challenge of Effective Long-term Thinking in the UK Government and the Critical Role of Philanthropy

by
Lord Robert Kerslake
*Former Head of Home Civil Service and Chair, Peabody Trust*
and
Christine Wagg
*Archivist, Peabody Trust*

*This chapter looks at the challenges to effective long-term thinking in government and the efforts that have been made to overcome them. It explores how a single philanthropist, George Peabody, was able to pioneer a brand of practical long-term thinking that has made a lasting impact in tackling poor housing and poverty in London. While success in tackling long-term challenges such as inequality and climate change depends on having good government, Peabody's example shows that philanthropy has a vital part to play. To the extent that the charitable sector outperforms government in its long-term vision, it can fill in the gaps left by government action. Charity can also lead by example, taking large philanthropic risks to provide proof in principle for its aims. And it can show government how to evade the grip of the "political business cycle," by creating independent bodies insulated from the troubles of the present day, by employing long-term commitment mechanisms to ensure that long-term plans are not abandoned in times of great urgency, and by placing small bets on promising policy proposals to see whether they bear out, and then scaling these proposals when the they prove to be successful.*

How good is central government at long-term thinking? This is a question that has long occupied the minds of senior civil servants, if not their ministers. Successive efforts have been made over a long period to strengthen the government's capacity and capability to undertake long-term work, with some notable individual successes. Taken in the round though, it still represents an area of significant systemic weakness

as the recent challenges of responding to the coronavirus pandemic have demonstrated. Even where good long-term analysis has been made, it has rarely been translated into action.

If the public sector struggles with the challenge of long-term thinking, the incentives for the private sector to do this are, if anything, even less. Much has been written about the distorting effect of the undue focus on annual profits and shareholder returns, so there is no need for me to repeat this here.[1]

Given these challenges to long-term thinking in both the public and private sector, it often falls to the charitable and community sectors to take the lead on this.

In this chapter I want to look at the challenges to effective long-term thinking and the efforts that have been made to overcome them. I will also explore how a single philanthropist, George Peabody, was able to pioneer a brand of practical long-term thinking that has made a lasting impact in tackling poor housing and poverty in London.

While success in tackling long-term challenges such as inequality and climate change depends on having good government, philanthropy (and the charitable sector more broadly) has a vital part to play. We should properly recognise its value as a resource when we are planning for the future.

## 1. The Problem of Long-term Planning in Government

I should start my critique of the government's long-term planning by saying first how much there is to admire in our civil service. At its best it combines great intellect, skill and agility in serving the government of the day.

The UK has an impartial but not independent civil service. Contrary to most perceptions, the vast bulk of civil servants are not involved in giving advice to ministers but in the delivery of public services such as providing employment support and benefits, running prisons and

collecting taxes. Total numbers of civil servants fell quite sharply during the period of austerity under David Cameron's coalition government and only began to rise again under Theresa May as the country geared up to leave the European Union.[2]

The combined effect of a constrained capacity and a rightful focus on the needs of ministers lies at the heart of why the civil service struggles to sustain the necessary focus on long-term thinking.

I have never met a minister who isn't a minister in a hurry. They recognise that their government is on a short electoral timescale – five years at most – and that they themselves are likely to be on an even shorter timescale in their current ministerial role. There is, in short, an enormous pressure to deliver on the demands of the day and deliver on their manifesto commitments. This gives their work the urgency and focus needed to overcome immediate obstacles, but it also makes them much less invested in long-term considerations. Long-term targets may be set with the best intentions, for example on reducing child poverty, but the focus is almost always on responding to the here and now. The civil service recognises this and deploys its resources accordingly. Any resources for future thinking have to be scraped from what is left over after the immediate demands of government are met.

There is an added challenge here in the centralised nature of government in the UK. According to the OECD, we are one of the most centralised countries in the developed world, with powers and funding much less devolved.[3] Ironically, according to work by the UK2070 Commission which I chair, we are also one of the most regionally unequal.[4] This centralising of power results in an overburdened centre, with far more of the decisions that might be taken locally elsewhere ending up on the desks of ministers and civil servants. This in turn adds a further squeeze on the capacity for long-term strategic thinking.

It is important to recognise the initiatives that have been taken in

government to develop long-term thinking, most notably, the Government Office for Science's Futures team, Horizon Scanning Programme team and Foresight project teams.

Foresight project teams work across government departments and with experts and academics to identify where new or emerging science can inform policy. The projects need to have a strong science or research element. Last year was the twenty-fifth anniversary of the publication of Foresight's first panel reports. In that time Foresight projects have informed government strategies on a wide range of issues including ageing, the seas, obesity, flooding and wellbeing. A number of these have been particularly influential.

Taken in the round, though, the impact on government has been at best uneven. At times, particularly when Sir David King was the chief scientific advisor, hugely influential. At other times, much less so. A good example of this was the work on flooding, published in 2004, but investment in flood prevention was only increased belatedly after many further instances of severe flooding.[5] The importance of thinking about the future is in theory recognised by the government as fundamental to policymaking. In reality it is much less clear. The effect of the Foresight work has been at best sporadic rather than embedded in government processes.

While there are challenges in the way that the government approaches the long term, the charitable sector has often performed much better. As I aim to show, the relative successes of the charitable sector have much to teach government actors.

**2. Lessons from the Charitable Sector: The Story of George Peabody**

A striking example of what it is possible to achieve is found in the life and work of the London-based American merchant banker George Peabody (1795-1869). He has justly been termed 'the father of modern philanthropy'. His benefactions were aimed at good causes that lay outside

the scope of both American and British public authorities at the time. Their total value exceeded £1.7m – a vast sum in the mid nineteenth century.

Peabody can be seen as the first major American philanthropist. In his birth country he founded initiatives that were centred on education, museums, libraries and scientific research, while in his adopted city of London he decided to tackle the desperate shortage of healthy and affordable working-class housing.

The third child in a family of eight, George Peabody was raised in modest circumstances in a small town in Massachusetts. He left school at the age of eleven to begin an apprenticeship in a general store; despite his later wealth and success he always regretted his meagre schooling. In 1831 he wrote: *'Willingly would I now give twenty times the expense attending a good education could I possess it, but it is now too late for me to learn and I can only do to those that come under my care as I could have wished circumstances had permitted others to have done by me,'*[5] When he founded the first Peabody Institute in his home town in 1852 he provided it with a motto – *'Education: a debt due from present to future generations.'* He achieved much in his long career, but the brevity of his school days clearly influenced his approach once he was able to give away large sums of money to benefit others.

From an early age he vowed to become a philanthropist if he prospered, and through diligence and enterprise he rapidly established himself as a successful businessman. His initial letter to his London trustees in 1862 included words outlining his objective: *'… from a comparatively early period of my commercial life I had resolved in my own mind that, should my labours be blessed with success, I would devote a portion of the property thus acquired to promote the intellectual, moral, and physical welfare and comfort of my fellow-men, wherever, from circumstances or location, their claims upon me would be the strongest.'*[6]

He went into partnership with Elisha Riggs who was fifteen years his

senior but had been impressed by the younger man's abilities. Operating first from Georgetown and later from Baltimore they bought and sold goods from merchants who imported much of their stock from England and continental Europe. George Peabody saw opportunities to negotiate the sale of American cotton in Lancashire and to buy wool, linen and other dry goods from Britain and ship them to America. Between 1827 and 1837 he made five trips to Europe, steadily increasing the volume of business conducted by his firm. Before crossing the Atlantic for the first time he made a will; out of a total of $85,000, about one twentieth was to go to good causes. He made a second will in 1832 and this time the proportion of his growing wealth intended for philanthropic purposes had increased to a quarter.

In 1838 he opened an office in the City of London and from that date he made London his permanent home, eventually giving up the dry-goods business and becoming a full-time investment banker. America was undergoing vigorous westward expansion and funding was needed for the construction of new railroads and canals. Peabody invested heavily in several such ventures and also assisted individual States when they were raising capital through the issue of bonds. His office in Moorgate became a meeting place for visiting American businessmen and he did much to strengthen Anglo-American relations. When the Great Exhibition took place in Hyde Park in 1851, he lent money to facilitate the USA's display of inventions and manufactured goods. He supported a project to lay a cable under the Atlantic to link Britain and America.

### 3. The London Housing Crisis and the Peabody Trust

By the middle of the nineteenth century London was expanding rapidly, but living accommodation was wholly inadequate for a rising population and the situation was made worse by the clearance of large areas to enable the building of railway termini. The public authorities took no responsibility

for the provision of housing; unscrupulous private landlords charged extortionate rents for grossly overcrowded and unhealthy dwellings. Local government was ill-equipped to deal with the impact of the capital's growth – it was in the hands of several hundred *'vestries'*, each covering only a small area. Their principal responsibilities were streets, sewers, firefighting and open spaces. In 1855 the first London-wide authority, the Metropolitan Board of Works (MBW), was set up, but its powers and funding were limited and it was appointed rather than elected, so the board lacked accountability and was generally unpopular.

In the years that followed, Peabody decided that he needed to take responsibility for London's housing crisis. The founding of the Peabody Trust was announced in the British press in March 1862, with its mission to *'ameliorate the condition of the poor and needy of this great metropolis,'* and it received widespread acclaim. Its first estate opened in Spitalfields in 1864 and within six years others were completed at Islington, Shadwell, Westminster and Chelsea. They were all designed by Henry Astley Darbishire (1825–1899) who served as the trust's sole architect for the next quarter-century. Although the flats were not self-contained – tenants had to share water closets and sinks with neighbours – they were a vast improvement on the common lodging houses or slum tenements which had previously been the only option for many workers and their families. Rents were affordable, communal laundries were provided and the blocks of flats were grouped around courtyards to provide safe playing areas for children. At each estate a resident superintendent collected the rent, handled applications for new tenancies, inspected the shared sanitary facilities for cleanliness and generally kept good order.

The early efforts of George Peabody's trustees met his approval, and in 1866 he gave them another £100,000. When writing to them about this additional donation he stressed his wish for the trust to have long-term benefits for London and Londoners. He said:

*'... it will form a fund the operation of which is intended to be progressive in its usefulness as applied to the relief of the poor of London ... without exclusion in consequence of religious belief or political bias. It will therefore act more powerfully in future generations than in the present; it is intended to endure forever. A century in the history of London is but a brief period comparatively with the life of man, and should your successors continue the management of the charity as you have begun it, it is my ardent hope and trust that within that period the annual receipts from rents for buildings of this improved class may present such a return that there may not be a poor working man of good character in London who could not obtain comfortable and healthful lodgings for himself and his family at a cost within his means.'*[7]

Peabody was not the first person to show concern about the capital's housing shortage. Since the early 1840s a number of philanthropists had attempted to improve living conditions for Londoners on low incomes, Angela Burdett-Coutts and Lord Shaftesbury being among the pioneers. However, the housing of the working poor was such a vast problem that what was achieved before 1862 had been on too small a scale to have a noticeable effect. George Peabody's gift was different.

The sum he gave was large enough to provide a significant number of dwellings from the outset, and it was an outright gift rather than a loan which would ultimately have to be repaid, or yield an income for the benefit of either the original donor or a group of shareholders. Other organisations usually aimed for a five per cent return on the original capital – hence the widely-used expression *'five per cent philanthropy.'* Peabody's trustees resolved at an early stage to fulfil the founder's wishes by making the fund self-perpetuating, so that its benefits could be spread over several generations. George Peabody approved their proposal that the trust's capital should produce a three per cent return – an impact investment that enabled the trust's long-term growth, but which was

sufficiently far below commercial rates to ensure affordable housing prices.

It was George Peabody's standard practice with all his benefactions to devote a good deal of time and preparatory thought to what he wished to achieve and how this could best be done. He selected his trustees with great care and began most new ventures with a fairly modest, exploratory donation. Once he felt confident that the trustees were moving in the right direction he would often increase the original sum to enable them to expand their work. Although he aimed to interfere as little as possible, he kept a close eye on progress. The itineraries for his occasional trips to America would include stops at places which had benefited from his generosity. In London he paid a personal visit to the second completed Peabody estate and expressed his satisfaction at the accommodation provided for residents.

George Peabody's views were quite advanced for his time; he regularly stressed his wish to exclude both political and religious influence from the management of organisations that bore his name. The directions he gave to his London trustees stated that these principles should govern both the appointment of the board members and the selection of applicants for accommodation. When he founded the Peabody Education Fund in the USA – aimed at healing many of the divisions caused by the American Civil War – the money was intended to benefit schools in the southern states and improve the training of teachers. He stipulated that the fund was to benefit all races; this was at a date when the abolition of slavery throughout the USA had only just become law and was still fiercely opposed by many white Americans.

### 4. Peabody's Long-term Legacy

Peabody died in November of 1869, but this did not stop his trust's progress. By the mid 1870s there was a growing recognition of the need for legislation to permit slum clearance schemes in some of London's

worst areas. The resulting Acts of Parliament gave the Metropolitan Board of Works the first powers of compulsory purchase. The sites it acquired with these new powers were offered to organisations or individuals willing to build 'model dwellings' to rehouse the slum tenants. Thanks to its financial strength, the Peabody Trust was able to buy more of the available sites than any other entity working with the MBW. The result was the construction of eight new estates and the provision of additional blocks at two existing ones. By the time Henry Darbishire retired he had designed nineteen Peabody estates, providing over 5,000 flats and housing more than 20,300 people.

In his will, George Peabody left yet more money to the Peabody Trust, making a total of half a million pounds given to provide affordable housing for Londoners. The present-day equivalent is roughly £55m.

The example set by Peabody was followed by many other philanthropists. Notable instances in the USA included the millionaires Cornelius Vanderbilt, John D. Rockefeller and Andrew Carnegie. In England the early success of the estates built with Peabody's money inspired Edward Guinness, William Sutton and Samuel Lewis to set up similar housing trusts using their personal wealth.

George Peabody had an even closer influence over a friend and contemporary who – like George himself – never married or had children but made a large fortune from a successful business career. This man was Johns Hopkins (1795–1873) who, acting on advice from Peabody, left more than $7m in his will to found institutions in Baltimore which are today world-renowned, including a library, a university and a hospital. A friend of both men later recalled a conversation in 1856 in which Peabody told Hopkins:

> *'When aches and pains made me realise that I was not immortal, I felt, after taking care of my relatives, great anxiety to place the millions I had accumulated, so as to accomplish the greatest good for humanity. I … formed the conclusion that there were men who were just as anxious to work with*

*integrity for the ... suffering and the struggling poor as I had been to gather fortune ... I gathered a number of my friends ... and proposed that they should act as my trustees. I then for the first time felt there was a higher pleasure and a greater happiness than accumulating money, and that was derived from giving it for good and humane purposes.'*

Over 150 years after its foundation, the Peabody Trust, now simply known as Peabody, continues its mission to help people make the most of their lives. It is responsible for 66,000 homes across London and the south-east of England where 133,000 residents live. Its housing stock includes many new developments and today there is an emphasis on building high-quality homes which integrate seamlessly with their neighbourhoods. The redevelopment of St John's Hill estate in Battersea and the long-term regeneration of Thamesmead in south-east London are just two examples. Properties have also been added to the portfolio through mergers with other housing providers. Sales of new dwellings at market price help to finance affordable housing; over 40 per cent of the homes built by Peabody in 2019–2020 were for social rent.

The charity's activities have expanded far beyond the provision of housing and now cover a range of social purposes. Over 18,000 vulnerable people are helped through its care and support services and during 2019–2020 it invested £9m in community activities across the Peabody Group. It also assisted over 1,100 people to find jobs and apprenticeships.

While Peabody has needed to grow and adapt to an ever-changing social and economic landscape, it remains true to its founder's aspirations to *'ameliorate the condition of the poor and needy ... and to promote their comfort and happiness.'* His foresight in setting up the Peabody Trust, his direction to his trustees that it should be self-perpetuating, and above all his generosity in giving away such a large sum of money, have ensured that the organisation continues to thrive. In George Peabody's own words, *'It is intended to endure forever'*

## 5. Conclusions: Long-term Planning and the Role of Philanthropy

I started this piece with a critique of government planning and the idea that the government might learn much from philanthropy's relative success at long-term planning. So what might the government learn from the charitable sector generally, and from the example of George Peabody specifically?

In some ways, it is perhaps inevitable that the third sector outperforms the first and second sector in its orientation towards the long term. Where the charitable sector is largely constituted by relatively impartial, altruistic ambitions, the incentives of governments tend towards benefiting the governed, and the incentives of markets towards benefiting buyers and sellers. In this way, the charitable sector is unconstrained by incentives to aid those who live today rather than the many who will live on our earth in the future, and is more easily able to adopt a long view.

To the extent that the charitable sector necessarily outperforms government in its long-term vision, it is able to fill in the gaps left by government action. Here government can support philanthropic activity and help it to flourish, so that it can continue to fulfil its pivotal role. Charity can also lead by example, taking large philanthropic risks to provide proof in principle for its aims, in order to create the political will for government to then follow its lead – just as Peabody demonstrated the possibility of affordable housing before governments had the capacity to take ownership of the crisis.

But government is not helplessly at the mercy of short-term political incentives, and to the extent that it can learn from charity to improve its responsiveness to the needs of future generations, it ought to do so. The public sector can seek independence from the 'political business cycle,' by setting up institutes insulated from the political will and which can therefore adopt a long view. It can follow Peabody in employing trusts with extremely long time horizons, and by adopting other long-

term commitment mechanisms to ensure that long-term plans are not abandoned in times of great urgency. Finally, government should adopt Peabody's venture capital model of risk-taking, placing small bets on unusual but promising policy proposals to see whether they bear out, and then following through with large sums when the proposals prove to be successful.

If we have the sagacity to learn from the charitable sector and its many luminaries, including the great George Peabody, then we may reasonably hope that the fruits of our own labour, too, will manage to endure forever.

CHAPTER SEVEN

# A Little Bit of Funding Goes a Long Way: The APPG for Future Generations

by
Sam Hilton
*Co-coordinator, All-Party Parliamentary Group for Future Generations and Deputy Director at Charity Entrepreneurship, Research Affiliate, Centre for the Study of Existential Risk*

*This chapter, written in August 2020, tells the story of the founding of the All-Party Parliamentary Group for Future Generations in the UK and details its achievements. The APPG Future Generations was founded by two Cambridge students, Tildy Stokes and Natalie Jones, in Autumn 2017. With the work of roughly one year of a full-time staff equivalent, the APPG supported the creation of Lord Bird's Well-being of Future Generations Bill, launched an ongoing Inquiry on Longtermism in Policymaking, pushed the UK Parliament to set up a Select Committee on Risk Assessment and Risk Management and swelled membership to seventy-five UK parliamentarians. The APPG members list continues to grow, and it aims to have more than 10% of all parliamentarians in the APPG in 2021. A little bit of funding goes a long way: the APPG has ambitious plans to continue pushing for a world where decision-makers at all levels of government fairly consider the interests of all future generations and can effectively plan for the long term.*

It was 6:40pm on Monday the 22 January, 2018. It was a chilly day outside but I had made it through the long queue and was now warming up in an ornate room in the Houses of Parliament. The faces of long-dead politicians lined the wood-paneled walls and stared down at a crowd of about forty people who were packed into a room that should have held no more than thirty. At the front of the room, across a sea of expectant faces, two students from the University of Cambridge, Tildy Stokes and Natalie Jones, were announcing the launch of the All-Party Parliamentary Group for Future Generations.[1] The part of the memory that stands out

most vividly was the face of Tildy's mother Dinah. She was standing at the back with a look of intense pride as she watched her daughter stand up to speak and kickstart a project that would soon be making political waves.

Let's rewind a bit. Back in early 2017, members of two Cambridge student societies, the Future of Sentience Society and the Wilberforce Society, had written an academic paper titled Rights and Representation of Future Generations in United Kingdom Policymaking. They believed that the future matters a lot. They were concerned both with their future as young students in an increasingly short-termist society, and the future of all generations: our children, grandchildren and beyond. They made the case that if we want to build a prosperous future – or even if we just want to prevent a disastrous future arising from catastrophes brought on by climate change, war or emerging technologies – then we need to improve politics. We need a political system that cares about future generations and can plan long-term. The paper itself is quite a dry read, but it did make a few concrete policy suggestions.[2] One of those was for the creation of an All-Party Parliamentary Group for Future Generations.

An All-Party Parliamentary Group (or APPG) is a group of UK parliamentarians (MPs and Peers) that meet to discuss a particular topic. There are APPGs on all sorts of things, from American Football to Zoos. They have a formal status in parliament, but no formal powers or formal funding, and many of the more active ones are funded by charities or companies. The paper suggested that an APPG for Future Generations 'may be a useful stepping stone to eventual institutionalisation of intergenerational justice in Parliament'.

And so Tildy and Natalie, setting their sights beyond the degrees for which they were studying, decided to turn the idea into reality with support from the University of Cambridge's Centre for the Study of Existential Risk. They spent their summer break in 2017 talking to everyone they knew who worked in policy, seeking connections to Members of Parliament.

They reached out to me, then just a lowly civil servant, and I doubt I was much help. But they forged ahead and the APPG was officially registered at a meeting in October 2017. And that is how I found myself in that room for the launch in parliament on that dreary Monday evening.

**January 2018 to March 2019: The power of funding**

When I was asked to write this chapter, the idea was to provide readers with a case study of what can be achieved, with a bit of grit and funding, towards creating a more future-focused political system.

Of course, you can't just buy political change, at least not in the UK. But a choice bit of funding to the right person with the right mission can get the wheels of politics moving. I was impressed by what had been set up by Tildy and Natalie, but they had exams and academic deadlines on the horizon, and limited capacity. And so I offered encouragement and help for them to seek out a small grant for someone to run the APPG. We raised £28,000 through the Centre for Effective Altruism and the Berkley Existential Risk Initiative.

I didn't expect to be the person running the APPG, but it took a while for the money to be found and by the time they were hiring I was available to take on the role. And so there I was, in March 2019, wandering the corridors of power with a small grant to cover my salary and asking myself the questions I asked Tildy and Natalie a year ago: how to organise a bunch of parliamentarians interested in the long term to drive political change?

**March to April 2019: Making plans**

In March 2019 the APPG had twenty-three parliamentarians on the member list. The secretariat, Tildy, had good relationships with a number of parliamentarians including Bambos Charalambous MP (chair) and Lord Bird (co-chair). Events were organised every three months, and poorly

attended by parliamentarians (although external guests were plentiful).

The £28,000 was to be split between myself, Sam Hilton, working four days a week, and my colleague, Caroline Baylon, working one day a week. Not a lot of funds but it would do for getting started.

Our somewhat floridly-expressed mission was: *'To provide impartial education, support and advice to Parliamentarians to assist them in ensuring that the UK Government takes into account the rights of future generations and is effectively addressing existential and catastrophic risks.'*

And our theory of change was simple: Grow the APPG + Research policy + Campaign → Policy changes

We were ready to go. Parliament on the other hand had different priorities. Parliament was preoccupied with Brexit, Brexit and more Brexit. There was a minority government. Brexit views did not neatly cut across party lines and MPs were rebelling, switching sides or forming new parties left, right and center. Every vote from every MP mattered. 2019 was to see more government defeats in parliament than in any other year, the controversy of parliament being unlawfully prorogued, and a general election. It was in this context that we made our plans. In March and April we met with our officers and developed the three core pillars of our work:

*Pillar 1 – Growth: Events and building the profile and members list*

We knew that our APPG was essentially an unknown force in politics. We needed to establish its reputation in parliament, build its members list, run events that attracted parliamentarians, and get them talking about the future and how to make policy for the long term.

*Pillar 2 – Research: An inquiry into longtermism in UK policymaking with Bambos Charalambous MP*

Under the shade of carefully cultivated trees in the large, glassy entrance hall of Portcullis House (where MPs have their Westminster offices), we

met with Bambos Charalambous MP, the APPG chair.

Bambos was a new MP, having been elected for Labour in Enfield Southgate in 2017. He was, and still is, keen to make as much of a difference as possible. He is always curious and has wide-ranging interests across policy topics — including an interest in technology, science and the future to go with his love of science fiction. He was keen to understand how we might build a parliament that works for the long term.

To meet his interest, and our own need for direction, we would launch an inquiry into 'long-term policymaking'. We would pull together a steering committee of MPs, run events, bring expert speakers into parliament, learn where long-term policymaking works well and produce a report setting out how to orient policy to the future.

*Pillar 3 – Campaign: The Today for Tomorrow campaign for a Future Generations Act with Lord Bird*

We met with APPG co-chair Lord Bird and his team in a modern meeting room that was an extension to the 19th century buildings of Darwin College, in Lord Bird's home town of Cambridge.

In early life John Bird ended up homeless and wanted by the police for petty crimes. He eventually found work in publishing and went on to found The Big Issue magazine to help give work opportunities to the homeless. For these achievements he was made a Lord. He is ambitious: his aim is to eradicate poverty in the UK. To do this he believes the UK needs a less short-termist political system that can work preventatively and stop people from sliding into poverty.[3]

We discussed plans for a UK-wide Future Generations Bill to tackle political short-termism. Based on the existing model in Wales, this would mandate politicians and civil servants to factor the long term into their plans and put a Future Generations commissioner in place to oversee it. We would work closely with Lord Bird and his team to draft

a Future Generations Bill and push it over the parliamentary hurdles until it became law.

**April to November 2019: The Brexit Era**

Throughout 2019 we found that parliament (especially the commons) was distracted by the Brexit back-and-forth and lacked the time to stop and think about the fact that they never make time to stop and think about the future. We put on a few events which attracted a few attendees but didn't seem to generate huge interest.

And so a lot of the early days of the APPG was spent in research, writing and planning with the hope that this work would be useful later on. We carried out interviews for our inquiry, made connections with other aligned interest groups and worked with Lord Bird's office to draft a UK Future Generations Bill, which he then laid as a Private Member's bill in the House of Lords. We reached out to members of the House of Lords to build support for the bill.

I wouldn't exactly regard those few months as a huge success, but in November 2019 when the Brexit stalemate ended with a general election, we noted that both the Labour party and the Green party had added a 'Future Generations Bill' to their manifesto. Neither party came to power, but the direction we were pushing in was clearly of interest to some in Westminster.

**December 2019 to March 2020: Growing interest**

The new year was better. Brexit was now "getting done" and there was a new parliament with new MPs keen to find ways to make useful change. We hosted a number of events to kick off the new year including an AGM which saw more than fifty parliamentarians fill out a packed Committee Room, and a fancy reception to formally launch the campaign for the Future Generations Bill. We prepared a series of inquiry sessions in

which our members would interview experts about how to make policy work for the long term. In the two months following the election the number of MPs who were members of the APPG more than doubled (from sixteen to thirty-five).

Of course maintaining that level of momentum and interest would have been challenging even without COVID-19 hitting the country in March.

**March to June 2020: Coronavirus times**

The pandemic came and we adapted. We moved our meetings online. The campaign for a Future Generations Bill was put on hold as Lord Bird shifted his attention to making sure The Big Issue business, and the UK's homeless, survived the pandemic. We shifted our attention to policy related to future risks. We researched risk management policy in the UK. We got press attention for Lord Rees, one of our officers and early members, and founder of the Centre for the Study of Existential Risk. We successfully requested a House of Lords Select Committee on Risk Assessment and Risk Management.

**Measuring our success**

I have told you the story of who we are, what we did and what happened. But what does this mean and how can we measure it? In May 2020 we wrote up and quantified our efforts and impact:

From 05 March 2019 until 05 April 2020, running the APPG for Future Generations cost £38,000. Most of this went to funding one full-time staff equivalent.

We grew. Over those dates the number of parliamentarian members grew from twenty-three to seventy-five with a fair spread across the Lords and Commons and across political parties. Over that time we ran eleven events with increasing frequency and popularity.

We researched. The bulk of our research was fed into the Inquiry

on Longtermism in Policymaking, that has yet to be written up. But we also did shallow dives into emerging technology regulation, managing extreme risks, AI, autonomous weapons, biosecurity, COVID-19, climate change and nuclear weapons. We used this to support a range of actors with policy development, all of whom have given positive feedback on the usefulness of our support. Those we supported include civil servants working on AI policy and longtermist policy, academics at the Centre for the Study of Existential Risk at the University of Cambridge and the Future of Humanity Institute at the University of Oxford and other policy actors such as the School of International Futures and Alpenglow.[4]

We campaigned. The Future Generations Bill campaign has had traction with private members bills laid in both the Lords and the Commons; over seventy MPs and forty peers have expressed support for the campaign and there has been significant press attention. We also tried, unsuccessfully, to further influence manifestos pre-election. And as mentioned, we succeeded in pushing parliament to set up a Select Committee on Risk Assessment and Risk Management to make the UK better prepared for all risks, pandemic or otherwise, national or global, in light of COVID-19. And we have started campaigns that are yet to reach fruition such as pushing for a debate on refocusing security from a 'national security' framing to a 'human security' framing.

There are some signs that this work is contributing to changing where politicians focus their attention. As well as the aforementioned manifesto commitments, we have seen the use of the term 'future generations' increase roughly 2.5 times in parliamentary debates since the creation of the APPG, including a 40 per cent increase over the funded period.

**How does this make a better future?**

Policy influencing is a slow process that takes time and grit, and it sometimes just doesn't succeed. It needs people, and funders, who can

work flexibly to grasp opportunities as they appear and adapt plans to an ever-changing political climate, and yet combine this agility with consistent movement in the right direction.

Over the last year and a half the APPG for Future Generations has had one significant success: getting parliament to set up a Committee on Risk Management. This will be a year-long project where relevant experts from the House of Lords will research this topic, have the power to call government actors to give evidence, and make recommendations to the government. About 40 per cent of Select Committee recommendations are implemented, and given COVID-19 and the uncontroversial nature of this work, I expect the government to be relatively receptive.

Beyond that, the bulk of our work has been to build momentum. The waves that Tildy and Natalie started are making bigger waves. And where this all goes next we have to wait and see. The APPG members list has grown and continues to grow, and we hope that in 2021 we can have more than 10 per cent of all parliamentarians in the APPG. We have ambitious plans to keep pushing for a world where decision-makers at all levels of government fairly consider the interests of all future generations and can work and plan for the long term. And we are optimistic that, with our one full-time staff member and a little bit of funding, we will one day make this ideal a reality.

I am optimistic. Politicians across the political spectrum know that democracy is short-termist. And although they have many pressing issues to deal with, they recognise that short-termism needs addressing. Making a better politics that works for future generations has broad appeal, there is a huge potential to make change and an ever growing movement of actors working hard on the issue.

# III

# Problem Areas

CHAPTER EIGHT

# Biosecurity, Longtermism, and Global Catastrophic Biological Risks

by
Jaime Yassif
*Senior Fellow, Nuclear Threat Initiative Global Biological Policy and Programmes*

*The COVID-19 pandemic has demonstrated the devastating impacts that high-consequence biological events can have on human lives, economic wellbeing, and political stability. While national and global leaders are rightly focused on saving lives and fostering economic recovery, now is also the time to strengthen international capabilities to prevent and respond to future high-consequence biological events – which could match the impact of the current pandemic or cause damage that is much more severe. COVID-19 should prompt global leaders to take bold action to reshape international institutions and make significant investments to reduce future pandemics and globally catastrophic biological risks. In working toward these goals, we must maintain a broad perspective about the potential sources of such risks. While naturally emerging novel pathogens can cause significant harm, engineered or synthesised pathogens have the potential to pose even greater risks. To address these risks, the biosecurity community should work with longtermist communities to prevent catastrophic laboratory accidents with engineered pathogens, and to prevent the exploitation of the legitimate global life science and biotechnology enterprise by malicious actors. It will be equally important to address the root causes of potential future bioweapons development and use by states and other powerful actors – including by strengthening the capabilities of international institutions to prevent and deter these activities. Now is the moment to accelerate progress and build wider coalitions around the shared goals of reducing global catastrophic biological risks and building a safer world now and for the long term.*

The COVID-19 pandemic has demonstrated that high-consequence biological events can threaten human populations and human lives, while also undermining economic and political stability. It is not the first global pandemic the world has faced, and it won't be the last. We

can reasonably anticipate more severe events in the future – including those that could pose a catastrophic or potentially an existential risk to humanity. Global catastrophic biological risks (GCBRs)[1] are biological events that could have devastating, population-wide consequences. Existential risks (X-risks) are the most extreme risks, and are defined as ones which can seriously undermine the long-term potential of human civilisation or threaten its very survival.[2] With the current focus of global and national leaders on biological risks, now is an opportune moment to strengthen partnerships across the biosecurity and longtermism-focused communities to cement an emphasis on protecting future generations in prevention and response preparedness plans.

Unfortunately, the world is woefully unprepared to cope with severe global pandemics, and COVID-19 has highlighted many of the vulnerabilities long warned of by biosecurity and public health communities. While decision-makers are rightly focused on saving lives and fostering economic recovery, now is also the time to strengthen our capabilities to prevent and respond to future high-consequence biological events, which could be orders of magnitude more severe.

**1. COVID-19 as a Warning Shot**

We can think of the COVID-19 pandemic as a warning shot.[3] [4] [5] The global spread of the SARS-CoV-2 virus is causing the continued loss of human life around the world, along with economic disruption and political instability. However, this crisis also offers an opportunity for those working to reduce global catastrophic biological risks.[6] First, the COVID-19 pandemic expands the Overton Window[7] on high-consequence biological threats, by helping leaders understand that they are possible and even likely, and by concretely illustrating their devastating consequences. It is now much easier to talk about GCBRs than it was as recently as one year ago. There is greater understanding of how a biological event can

rapidly spread to every continent, with the potential to infect 40 to 70 per cent of the global population – while having a case fatality rate[8] exceeding that of a bad seasonal flu, and causing significant economic and political damage. Second, COVID-19 demonstrates the importance of proactive, flexible preparedness for novel and unanticipated pathogens – not only the need for rapid development of vaccines and therapeutics, but also for capacity to rapidly scale public health and non-pharmaceutical interventions in response to high-consequence biological events that could escalate into a GCBR-scale event. This type of preparedness gives leaders the necessary tools to save lives and prevent disease transmission during very large events that can overwhelm the capacity of conventional health systems. The bottom line: COVID-19 has exposed vulnerabilities and is creating a window of opportunity to make the case for sustained investment and focused work to build capacity to prevent, detect, and rapidly respond to GCBR-scale events.

The lessons of COVID-19 should prompt global leaders to take bold action to establish new financial instruments and reshape international institutions to reduce large-scale biological risks. There are now opportunities to drive significant institutional changes which can meaningfully reduce GCBRs – including outcomes that would have been much more difficult to achieve even a year ago. For example, the Nuclear Threat Initiative (NTI) recommends establishing a high-level coordinator in the United Nations[9] dedicated to bolstering preparedness for high-consequence biological events, including those that could reach a GCBR scale. This senior official would be responsible for staying abreast of emerging biological risks and iteratively testing and strengthening UN and World Health Organisation (WHO) capacity to marshal an effective, integrated response to biological events from a range of sources. It is also important for leaders at the highest levels of government to identify weaknesses in pandemic preparedness capacity and build specific fin-

ancing for reducing catastrophic biological risks into their budgets.

Finally, it's important for national and global leaders to adopt new strategies and develop operational capabilities for early detection and proactive response to high-consequence biological events and GCBRs. This includes adopting clearly-defined sets of triggers to drive early, anticipatory action which can quickly contain disease outbreaks that could otherwise escalate exponentially. This work should also include development of effective strategies and capabilities for deploying interventions at scale to reduce mortality, block chains of transmission, and gather actionable information about population-level transmission dynamics to guide an effective response.[10] Had these systems and tactics been in place for COVID-19, hundreds of thousands of lives might have been saved; it is important to start building these capabilities now to prepare for the next major global biological event. This type of anticipatory planning can help facilitate an important paradigm shift: it helps public health professionals and leaders conceptualise and develop the capacity for plan B when conventional health systems become overwhelmed and other approaches will likely be needed to save lives and prevent further disease transmission. This type of planning can also help bridge the divide between current operational capabilities, which are already straining to cope with COVID-19, and the tools and systems that the world would need to marshal an effective response to a much more severe GCBR-scale event.[11]

These changes will not come easily. Translating this crisis into impactful action will take steady attention, advocacy, and focus. Biosecurity-focused organisations and experts will need to engage in a sustained effort to ensure that international organisations and national leaders make the necessary institutional adjustments and sustained investments to empower them to prevent, detect and respond to GCBR-scale events.

Biosecurity leaders must continue to emphasise that COVID-19 is not

the first global pandemic we've faced, that it won't be the last, and that we can reasonably anticipate events in the future that are equally severe or potentially orders of magnitude worse. We must also remind public health, security, and national leaders that these events can originate from a variety of sources – not only naturally emerging novel infectious diseases, but also deliberate and accidental releases.

## 2. The Most Significant Sources of GCBRs

While naturally emerging novel pathogens can cause severe damage – as we are currently experiencing with SARS-CoV-2 – engineered or synthesised pathogens have the potential to pose an even greater risk of a biological event with devastating global consequences. We must maintain a broader perspective about the most significant potential sources of GCBRs and X-risks. This question can be broken down into two components: (1) the types of pathogens or biological agents that pose the greatest risks, and (2) the types of scenarios that could lead to a GCBR-scale event or an existential threat.

A commonly held view among longtermist communities is that states pose the greatest risk of creating a GCBR. The reasoning behind this argument is that states have the greatest available resources and access to talent, and so their capabilities are effectively unconstrained. In terms of intent, states are generally motivated by the rational pursuit of political, military, and economic goals – none of which are well served by deliberately causing a GCBR or existential bio-risk. However, perverse incentives can arise from competition for influence and resources within bureaucratic structures which can drive development of bioweapons even if national political leadership does not support that goal. It is also possible that states could be interested in developing a biological doomsday device as a means of deterrence in cases where nations cannot obtain a nuclear weapon or conventional means of deterrence.

At NTI, we understand the importance of reducing the likelihood that a state would deliberately or accidentally cause a GCBR, and we are working to address the root causes of potential bioweapons development and/or use by states and other powerful actors. Over the past two years, our work has included: developing actionable recommendations to reduce the potential for bioweapons programme proliferation; galvanising international support for actions to reduce GCBRs through a series of tabletop exercises at the Munich Security Conference and through high-level engagement with United Nations leaders; and developing new mechanisms to financially strengthen the Biological Weapons Convention (BWC), which embodies the global norm against bioweapons development and use.

To accurately assess risk, it also is important to avoid prematurely discounting non-state actors as a potential source of a GCBR or X-risk. First, it should not be assumed that sophisticated non-state actors would be unable to create a biological agent that could cause an event of this scale. Rapid, globally distributed technology advances are continuing to lower the barriers to the synthesis and engineering of pathogens and other biological agents – thereby enabling a wider range of actors to engage in this type of work.[12] Second, with respect to intent, it is not difficult to imagine a non-state actor with an apocalyptic mission that might be aligned with the pursuit of weapons capable of causing catastrophic damage to humanity. There is publicly available evidence that these types of groups do exist. For example, the Aum Shinrikyo cult, which is widely viewed as an apocalyptic group, pursued the development of chemical and biological weapons and made multiple failed attempts at launching large-scale chemical and biological attacks in Japan in the 1990s.[13] In aggregate, these considerations lead NTI to the view that work to reduce biological threats posed by non-state actors can also have significant value for reducing GCBRs and X-risk, and we are actively working with international partners to do so.

**3. Reducing GCBRs and Associated with Technology Advances**

As noted, emerging biological risks associated with rapid technology advances are a key driver of growing GCBRs and X-risk. To reduce these risks, the international community will need to fill several key gaps. First, it's important to establish a shared international perspective – or set of norms – for dual-use bioscience research: norms about how to determine whether dual-use research and development activities should move forward and how to weigh the perceived benefits of the work against the potential safety or security risks.[14] Second, it is vital to develop new approaches for governments, funders, the private sector, academic researchers,[15] and publishers to act on these norms. This includes developing and executing clear and effective governance mechanisms to oversee dual-use bioscience work from early-stage design and funding decisions, through project implementation, and on to publication.[16]

NTI is prioritising this work because it can help prevent the exploitation of the legitimate global life science and biotechnology enterprise by malicious actors. Our work includes catalysing new approaches to managing access to the goods and services needed to engineer and synthesise biological agents,[17] accelerating investments and innovation in biosecurity,[18] and preventing the public dissemination of information hazards.[19] If successful, these approaches could constrain the capabilities of malicious actors, including sophisticated non-state actors. While we appreciate that no single intervention in this arena will create a fail-safe protection against catastrophic biological risks, we are working to develop the critical elements of a layered defence that, in aggregate, significantly reduces them.

**4. Reducing Intent to Develop or Use Biological Weapons**

Reducing biological risks posed by states and other powerful actors is challenging work. One key element involves influencing the intentions of

these actors to shape their cost-benefit calculation. This means building stronger international capabilities to reduce dangerous misperceptions about bioscience and biotechnology activities in other countries, in order to avoid arms races. It is equally important to bolster deterrence and accountability for any violation of the international norm against the illicit development or use of biological weapons. To achieve these goals, it is critical to enhance transparency about bioscience and biotechnology research and development. This would help increase clarity and avoid unwarranted suspicions about the capabilities and intentions of legitimate programmes around the world. It is also important to develop more robust international capabilities for investigating the source of a biological event of unknown origin during a crisis.[20]

Currently, the WHO is empowered to conduct public health investigations, and the United Nations Secretary-General oversees a mechanism for investigating alleged state use of biological and chemical weapons. However, there is no mechanism for evaluating events that fall between these two poles, or when there is a lack of clarity or suspicion about the event source. This latter type of mechanism might have been useful in addressing recent accusations about the origins of COVID-19 – in an internationally credible, evidence based, and transparent manner. Finally, in the event of a clear violation of international norms, the international community should have stronger accountability measures in place so that states and other powerful actors have reason to believe that they would be held accountable for developing or using biological weapons.

## 5. Working Together to Reduce GCBRs

To achieve the ambitious agenda of preventing and responding to future pandemics and GCBRs, we need all hands on deck: focused and consistent effort by talented and energetic experts; creative thinking to develop innovative solutions to hard problems; political will; and

financial resources. It will be important to build bridges and common cause across communities – including between national and global leaders, the security sector, the conventional biosecurity community, and communities focused on longtermism and protecting future generations. This begs the question: how can longtermist experts and biosecurity partners achieve buy-in from more conventional leaders and decision makers on the goal of reducing GCBRs?[21]

First, it will be important to bridge these communities and build a mutual understanding that bio-risk reduction policies and operational response capabilities must focus not only on near-term crises but also on protecting future lives. NTI has been successful in bringing a focus on GCBRs to conventional meetings of foreign policy and national security decision-makers, including by convening an annual tabletop exercise at the Munich Security Conference and co-leading an event on GCBRs at the World Health Assembly. These and other efforts, such as the Clade X exercise by the Johns Hopkins University Centre for Health Security,[22] are invaluable and should be expanded.

Second, we should consider aligning arguments about future lives more explicitly with simultaneous and related discussions about the X-risk associated with climate change. For example, the 'extinction rebellion' language that is being used by the growing, global movement of climate change activists clearly demonstrates that there is concern around the world about human extinction risks and momentum behind the growing movement to prevent that future. There is an opportunity to forge connections between this widely recognised global effort and work to reduce catastrophic and potentially existential threats associated with pathogens and other biological agents.

The sustainable development community should also be recruited to the GCBR cause in light of aligned interests in protecting and improving the quality of future lives, and the concomitant desire to guard against

the risk that GCBRs could significantly set back progress on achieving sustainable development goals. In our work at NTI, we have found that to make progress on reducing specific global catastrophic biological risks associated with technology advances, it is important to align with the development community around common interests: technology advances can support economic growth, bolster public health and help defend against pandemics and GCBRs. However, they also pose a great risk – if deliberately or accidentally misused – of causing a catastrophe that undermines those very goals.

Third, more attention must be paid to building a diverse set of next-generation biosecurity leaders who are thinking both near-term and long-term. The Emerging Leaders in Biosecurity Initiative[23] and 80,000 Hours[24] are both doing valuable, impactful work in this area. More programmes along these lines, and with greater global reach, would be tremendously valuable in continuing to advance this goal.

Finally, a more geographically, racially, and ethnically diverse group of stakeholders and leaders from different policy backgrounds should be included in the development of catalytic actions to reduce GCBRs and X-Risks. Low-income countries and under-served populations have much to lose in the face of emerging potentially catastrophic risks, and they should be regularly included in work to reduce them. The inclusion of more diverse perspectives in these discussions is likely to lead to more effective and creative solutions, which are workable for both the developing and developed world. Moreover, fostering a more diverse international coalition in support of GCBR reduction work will pay dividends, as it has the potential to build political will and drive progress within multilateral fora, such as the Biological Weapons Convention and other parts of the UN system.

Recent efforts to bridge the biosecurity and longtermist communities have already begun to bear fruit, and this partnership is helping gain

traction for GCBR and X-risk-reduction goals among the broader national security, global public health, and development sectors. Now is the moment to accelerate progress and build wider coalitions around the shared goal of building a safer world now and for the long term.

CHAPTER NINE

# Utilising Insurance for Climate Risk Reduction in the UK

by
Lord Des Browne
*Vice-Chair, Nuclear Threat Initiative and Co-founder and Chair, Executive Board, European Leadership Network*

*The insurance sector has an abundance of expertise, money, and perspective which it might bring to the mitigation of and adaptation to climate change and other catastrophic risks. However, the unique characteristics of the insurance sector remain largely unleveraged as a policy tool. This chapter describes the advantages policymakers can gain by leveraging the capabilities of the insurance sector in risk management. The insurer's role as a financial shock absorber is relatively well appreciated, but their role as risk engineers is less so, and offers a largely untapped pool of expertise in risk modelling, pricing, prevention, and behavioural incentives. This chapter also characterises the current hurdles to cooperation between government and insurers, including the evolution of siloed institutions, and the systematic neglect of long-term risk. It then identifies actions for government, the insurance sector, civil society, and philanthropists to improve collaboration and increase their collective influence. Of particular emphasis is the role of philanthropic investment in subsidising insurance products to make them available in developing nations.*

In September 2015, while speaking before an audience at Lloyd's of London, *'the bedrock of the insurance industry'*, the Bank of England's Governor Mark Carney called climate change the *'Tragedy of the Horizon.'*

He explained: *'We don't need an army of actuaries to tell us that the catastrophic impacts of climate change will be felt beyond the traditional horizons of most actors, imposing a cost on future generations that the current generation has no direct incentive to fix. That means beyond the business cycle; the political cycle; and the horizons of technocratic authorities, like central banks, who are bound by their mandates … In other words, once climate change*

*becomes a defining issue for financial stability, it will already be too late.*'[1]

According to one recent study, climate-change-driven disasters accounted for 91 per cent of the 7,255 major disasters between 1998 and 2017.[2] In 2017, there were losses of around US$340 billion from extreme weather events alone, over half of which was uninsured. In the most vulnerable regions, for those least equipped to recover, losses were almost completely uninsured. Arguably, such disasters are already a defining issue for financial stability in the countries most exposed to climate change.

However, to Carney's point, it is apparent that the recent economic and humanitarian costs associated with this recognisable upward trend will be but a drop in the (much warmer) ocean when measured over centuries, rather than decades. Without intervention, climate change in the centuries ahead will be less a defining issue for individual economies, and more an existential issue for the global economy and human societies as we know them.

In 2019, four years on from his remarks at Lloyd's and this time addressing the UN's Climate Action Summit, Carney emphasised the vital role of insurance in smoothing the transition to a 1.5-degree world:

'*[The insurance] sector brings three things: expertise, money and perspective and those are all crucial in helping society adjust to the reality of that transition.*'[3]

Horizon scanning, long-termism, fostering resilience, and understanding and offsetting the impact of catastrophic risks are in the insurance industry's DNA.

This chapter highlights the need for those who set public policy, including the government, and those who seek to influence policy, including philanthropists, to understand and engage with the unique capabilities of the insurance sector in order to more effectively evaluate, mitigate, reduce vulnerability to, adapt to and build resilience against climate change. To this end, the chapter describes the advantages policymakers can gain by leveraging the accumulated capabilities of the insurance

sector in risk management, along with the current hurdles to cooperation between the government and insurers. It then identifies actions for the government, the insurance sector, civil society and philanthropists to improve collaboration and increase their collective influence.

**1. Missed Opportunities: A Matter of Mindset**

The management of risk is a shared responsibility. The National Risk Register sets out the *'likelihood and potential impact of a range of different risks that may directly affect the UK.'*[4] Increasingly, it is common in publications of this nature for authors to deploy exhortations to non-government entities, including business, to coordinate with the public policy community to prevent and mitigate identified risks. For example, the UK National Biological Security Strategy[5] states that *'preventing biological risks … is not something that (the government) can do alone.'*

An HM Treasury policy document from March 2020[6] emphasises the shared nature of risk management, characterising the UK government's role as insurer of last resort. It describes the purpose of government intervention in this respect as, in substantial part, helping improve the market for insurance and providing protection against risks where the private sector is unable to provide full insurance cover.

The unique contribution that insurance can make to modelling, assessing, and incentivising businesses' transitions to more resilient behaviours is thus appreciated within the UK Government, but it is not yet fully appreciated in relation to climate change. The Home Office's national strategy for countering terrorism, CONTEST,[7] recognises that *'the insurance industry has the potential to shape behaviour and improve safety, security and resilience', with the sector placed uniquely to 'better protect our economic infrastructure and to scale our ability to tackle terrorism.'*

The unique and advantageous characteristics of the insurance sector remain largely unleveraged as a policy tool and this bias extends even

to those who seek to influence public policy. For example, despite the Committee on Climate Change's eminent position of authority and leadership in the field of climate change adaptation and policy, nowhere in its encyclopaedic June report[8] does it mention insurance or the resilience benefits of risk transfer.

There have been several recent, positive examples of cooperation from which lessons can be drawn. These include the relative success of Pool Reinsurance Company Ltd and Flood Re Limited, each a joint initiative between insurers and government, and ongoing discussions between insurers and government about the failure of the market to provide pandemic insurance. Generally, however, there is a frustration in the insurance sector that the government remains a relatively siloed institution. In a recent CASS Business School webinar, this frustration was articulated by representatives of the insurance industry, who commented on the lack of consultation and appreciation of their experience identifying, modelling, managing and mitigating the already evident effects of climate change.

It is difficult to comprehensively explain why policymakers don't engage more routinely with the insurance industry on risk. The most obvious reason is that these respective institutions have evolved over many years as independently functioning and siloed from one another. There has been no obvious reason to change the status quo.

In *The Precipice*,[9] leading existential risk researcher Toby Ord identifies further reasons why such risks are neglected. Among them, he identifies two reasons revealed by political science.[10] First, that *'the attention of politicians and civil servants is frequently focused on the short-term.'* He argues that their thoughts and actions are set to respond to the election and news cycles, and consequently they find it difficult to turn their attention to a problem that won't strike for several election cycles. Second, Ord reports that having raised the topic of existential risk with senior politicians and civil servants, he has encountered a common reaction:

genuine deep concern paired with the feeling that addressing the greatest risks to humanity is *'above my pay grade.'*

Climate change risks prove a major challenge because they are not suited to our ingrained treatment of short-term and conventional risks. Both their scale and gravity is overwhelming and they therefore lose the competition for the attention of politicians and civil servants to comparatively trivial problems of immediate salience. A new paradigm of risk management is required.

**2. Insurance as a Policy Tool: Money, Expertise, Perspective**

Insurers are risk managers protecting society's assets, but they are also long-term investors funding the economy, and they are influential in enabling and incentivising their private, corporate and government policy holders to change their behaviours. Given its money, expertise and perspective, the insurance industry is uniquely positioned to make substantial contributions to society's understanding and management of climate change risk.

*Money*

A primary purpose of the insurance industry in providing capital to support response and recovery is the provision of financial resilience.

*'Insurance, when put in place, provides financial resilience … it provides the flow of capital to support communities and infrastructure to recover from disasters. Without adequate insurance, the burden of paying for losses falls largely on individual citizens, governments or aid organisations, with significant impact upon already straining government budgets, and economic and social hardship for those affected … Countries with high insurance cover recover faster from disasters, and increasingly, governments are recognising the role and benefits of insurance in transferring risk from disasters. Yet there is a large and even widening "protection gap" of underinsurance.'*[11]

The phrase '*protection gap*' is used to describe the difference between losses which were compensated by insurance (protected) and those which were not (unprotected). Estimates put uninsured losses from natural catastrophes globally at 70 per cent, a protection gap which has not meaningfully narrowed in decades. One of the biggest mistaken assumptions is that meaningful protection gaps are only a problem for developing countries, an assumption that is now less likely to be made in the wake of COVID-19.

Recent studies[12] have provided evidence that countries with widespread market-based insurance coverage recover more quickly from the financial impacts of catastrophe. These studies repeatedly demonstrate that pre-emptive investment in defences and adaptations is several times more financially efficient than traditional post-disaster financial assistance. In fact, post-disaster financial assistance often serves to disincentivise people, businesses and local governments from taking proactive mitigatory action, leaving the underlying vulnerability unchanged.

As an indication of the payoff for countries able to lower their protection gaps, Lloyd's of London estimates that a 1 per cent rise in insurance penetration can translate to a 13 per cent reduction in uninsured losses and over 20 per cent reduction in the disaster recovery burden on taxpayers.[13] Substantial macroeconomic benefits include increased investment and higher output – potentially up to 2 per cent of GDP. Finally, on the asset side of their balance sheets, insurers can be highly influential in bringing the realities of climate change into mainstream financial decision-making. Estimates of assets under management by the industry vary, but are in the multiple trillions, and environmental, social, and corporate governance investment strategies are emerging as the predominant methodology of insurers. An example of such a strategy may be resisting investment in companies with more than 30 per cent of their business associated with thermal coal mining or coal power generation.

*Expertise*

While insurers' role as financial shock absorbers is relatively well appreciated, their role as risk engineers is less so. This is probably a significant factor in the industry being largely overlooked as a climate policy tool.

*'The insurance industry's potential … is grounded in its near real-time vantage point of risk insight. Every day, countless companies have their risks underwritten and receive payments for claims; behind the scenes, the insurance industry churns and analyses data to understand which risk management practices are working and which are not. For the segments of risk where this potential has been realised, top insurers are considered risk engineering advisers that happen to provide insurance, rather than the other way around.'*[14]

In national disaster-preparedness, insurers are a largely untapped pool of expertise in risk modelling, pricing, preventative measures and behavioural incentives to build socio-economic resilience to climate risks. Rigorous modelling, quantification and understanding of actual and potential catastrophic events are core competencies of insurance since they are necessary to generate an insurance product. These competencies enable insurers to incentivise and reward businesses and individuals who comply with accredited risk-reduction behaviours with a reduced policy premium.

However, there is a notable discrepancy in the fact that an insurance company may have the data to indicate the likelihood of loss – allowing it to build understanding of how to prepare for and mitigate the consequences of a disaster – but does not have the legislative power to enforce active preparedness steps, nor the influence over necessary regulation.[15]

This disparity is well-documented by the insurance sector and by academia. One frequent suggestion is the improved use of insurers' data and knowledge by national and local government in land-use planning and in developing zoning and building code regulations, standards, and construction requirements. Each of these elements are fundamental to

adapting to and mitigating natural disaster damage. Insurers' expertise in risk appraisal makes them uniquely qualified to support the most cost-effective preventative and *'Build Back Better'* initiatives. As noted by one recent report, *'investment must be directed effectively, especially in the context of climate change adaptation. Expenditures that seem worthwhile over a 10 or 20-year perspective may not be justifiable over a longer horizon.'*[16]

*Perspective*

Related to the insurance industry's expertise as risk engineering advisers is the perspective they are able to bring to the determination of where government involvement in the provision of insurance protection is necessary. Insurance works on the basis that the premiums of the many pay for the claims of the few. Given that so many people would be affected by them, the tail risks of long-term climate change are simply uneconomic for insurers to underwrite. Commercial insurance was not designed to be able to provide unlimited cover for risks which correlate nationally or even globally, as the COVID-19 crisis has demonstrated.

The limits of the commercial market for such risks leaves government, and by extension the taxpayer, as the de facto 'insurer of last resort' when catastrophes occur. There is a large protection gap which only public finances can fill. This responsibility of government has been articulated specifically in the HM Treasury's Government as insurer of last resort: managing contingent liabilities in the public sector.[17] Contingent liabilities are risks the government *'takes on that others cannot, both to protect the population and provide stability when unforeseen events occur,'* and are *'an increasingly important policy tool to support economic growth and safeguard ... the long-term sustainability of public finances.'* The National Risk Register lists at least ten '*natural hazards*' clearly linked to climate change which are considered contingent liabilities for which the government is insurer of last resort.

The document makes a series of welcome recognitions and recommendations about the importance of government providing guarantees to enable commercial insurance markets to develop around catastrophic risks to society. It uses the example of Pool Re, the UK's state-backed mutual provider of terrorism insurance, to note that:

*'By taking on the tail-end risk of a catastrophic event from the private sector, the government makes it possible for private insurers to re-enter the market … The Pool Re model has been successful in maximising the involvement of private insurance to enter the market, creating several layers of defence before the guarantee is called … These layers of defence have created a £10 billion buffer between an incident occurring and taxpayer money being called on.'*

The Pool Re partnership represents an effective but underused method by which government can positively intervene in a market by partnering with insurers to build financial resilience to a complex, catastrophic risk. The partnership also fulfils the 'expertise' potential of the industry by leveraging its capabilities and amplifying them through protective security initiatives with academia, the Metropolitan Police and the Home Office. In the context of a coordinated national climate resilience strategy, the merits of this multi-stakeholder approach should be considered when the government undertakes to 'consider its stock of contingent liabilities and investigate where it may be appropriate to expand the scope of current pooling schemes', as recommended by the HMT policy document.

### 3. What Role Can Philanthropic Investment Play?

It is critical that civil society, a third pillar distinct from government and business, is actively engaged with climate change policy. A particular responsibility of civil society is to ensure that national policymaking does not neglect protection of the poor and vulnerable. This is especially crucial in insurance policy, which is complex and, in its implementation, can be damagingly discriminating. As Toby Ord notes in *The Precipice*:

*'When citizens are empathetic and altruistic, identifying with the plight of others … they can be enlivened with the passion and determination needed to hold their leaders to account.'*[18]

Philanthropy involves the deployment of private financial resources (or time or initiatives) for the public good. There are leveraged synergies between philanthropic entities and the insurance sector, albeit of differing motivations and scales, in terms of money, expertise and perspective. This is the case particularly within the humanitarian and development aid context. Money can be directed toward building linked capabilities between government and insurers and toward making products more available and accessible to lower-income and developing nations.

Investment in research to improve open-access climate risk data and models can remove barriers to such products, which are otherwise expensive to develop and often protected under proprietary or commercial licenses. For example, a consortium of philanthropic organisations recently supported the successful development of an open-source platform for climate risk modelling and risk-financing decision-making by the public sector and (re-)insurers in the Philippines and Bangladesh.[19]

Subsidisation of insurance products can help with bridging the insurance protection gap, where such products are beyond the means of the vulnerable to purchase. A recent study[20] describes the successful scale-up of a livestock insurance programme in Kenya,[21] for which there was insufficient demand initially, following subsidisation of the farmers' premiums.

Many charities have expertise in humanitarian and development aid. Their expertise can be leveraged in the design of innovative insurance products for those in lower-income and developing nations.

Micro-insurance is one such example. Similar to micro-finance, these innovative products can be used to insure vulnerable individuals. The World Food Programme and Oxfam partnered to successfully implement the R4 Rural Resilience Initiative in Africa, which integrates risk transfer

with risk reduction to increase the resilience of the poorest farmers to specific climate risks.[22]

Parametric insurance products provide a timeliness and purpose flexibility well-suited to innovative applications for risk financing against climate-related disasters. In 2015, for example, Mauritania received a rapid payout of US$6m, triggered by drought metrics under the African Risk Capacity (ARC) programme, which successfully provided liquidity sufficient to mitigate a humanitarian crisis.[23]

The innovation of risk sharing partnerships, such as multi-sovereign pools, can support the insurability of nations unable to afford insurance products in the private market. For instance, the Caribbean Catastrophe Risk Insurance Facility (CCRIF) was a pioneer of such partnerships and *'has made successful disaster liquidity payments on a range of perils over the 10+ years since its inception, most notably paying more than US$61m during the 2017 year of hurricanes that affected the Caribbean.'*[24]

Despite such successes, these products have low penetration relative to their global potential. In the case of micro-insurance, this is due to a lack of tailored products, distribution channels and affordability. In the case of parametric and risk-sharing products, the limits come from challenges surrounding financial literacy and quality of data and models.

Philanthropic entities offer a unique perspective, stemming from altruistic motivation. Several longtermist organisations, such as Longview Philanthropy, use evidence, reason and analysis to identify the most effective ways to do the most good. The philosophical movement that guides such organisations prioritises a long-term perspective in efforts to address existential risks such as climate change. Such a perspective is powerful in building risk literacy and influencing discourse in society. Promotion of and investment in risk literacy can increase insurance penetration and effectiveness by enabling society to identify risks, analyse response options and allocate ownership of risk management.

Furthermore, a recent CASS study highlights that a key *'challenge with all insurance products is to improve their interaction with resilience measures in an integrated climate adaptation strategy,'* which should ensure that 'specific insurance solutions for financial protection also incentivise risk reduction and preparedness'.[25] In this vein, the philanthropic perspective can also help strengthen the dialogue around initiatives, such as *Build Back Better*,[26] with long-term resilience benefits.

The UK Government demonstrated early success in merging the political, business and philanthropic agendas. Between 2011 and 2015, the Department for International Development placed resilience to natural disasters at the heart of its development programmes, successfully supporting and implementing pioneering insurance tools to manage risk in countries across Africa and the Caribbean. With only a small fraction of its aid budget, the UK was able to set an international precedent that has since been adopted by other governments and development bodies.

Since then, however, political attention to these promising insurance-centric approaches to humanitarian and development aid has dwindled. As such, long-term philanthropic investment can play an important role in building capacity and ensuring transparency of engagement between governments and insurers to increase the resilience and adaptability of those vulnerable to climate change.

## 4. Conclusions: The Shift to a Long-term Perspective

Developing long-term domestic policy while responding to global climate change remains a formidable challenge. Securing our society's future will require breaking political and economic reward cycles that incentivise short-termism at the expense of integrated long-term action and horizon scanning. Despite the fundamental psychological barriers that must be overcome, there is emerging evidence of willingness to shift towards a long-term perspective in UK policymaking with the formation of The

All-Party Parliamentary Group for Future Generations,[27] whose role is to *'raise awareness of long-term issues, explore ways to internalise longer-term considerations into decision-making processes, and create space for cross-party dialogue on combating short-termism in policymaking.'*

This direction of travel was reinforced by the recommendation of the House of Lords Liaison Committee *'that a special inquiry committee be appointed "to consider risk assessment and risk planning in the context of disruptive national hazards" to report by the end of November 2021.'*[28] This decision was made in response to a proposal by Lord Rees of Ludlow that a special inquiry committee be set up to consider risk assessment and risk planning given that *'the United Kingdom is at risk from major disruptive hazards which have the potential to cause significant human, economic, environmental and infrastructure damage.'*

We are at the early stages of the realisation that the deployment of insurance as a policy tool can be effective in evaluating, mitigating, reducing vulnerability to, adapting to and building resilience against climate change.

The UK insurance market is world leading. Industry, government and civil society are natural allies. An alliance among them in respect of climate risks is a policy tool that we cannot afford to ignore.

CHAPTER TEN

# Ensuring the Safety of Artificial Intelligence

by
Amanda Askell[1]
*Research Scientist, Anthropic*[2]

*Technological innovation has been a cause of both great flourishing and great risk in the history of humanity. Artificial intelligence stands out among other forms of current technology as having particularly great promise and great risk given that it could one day perform every kind of intellectual work. This chapter makes the case that our work today to ensure the safety of artificial intelligence could have a significant long-term impact on humanity, and we should therefore begin this work now. It begins by explaining in historical context why progress on AI is likely to have a significant effect on the long-term future. Though the trajectory of this progress and the nature of its effects are uncertain, we can act today to meaningfully alter these for the better. We can do so by gathering information about AI progress and impacts, investing resources into the safe development and responsible deployment of AI, and working to resolve collective action problems that threaten to undermine these efforts.*

Technological innovation has been a key driver of human flourishing. The Green Revolution prevented the starvation of millions or even billions of people.[3] However, technological innovation has also been a key driver of new and often unanticipated risks. In perhaps the most salient example of this, the development of nuclear weapons played a crucial part in the risks to humanity presented by the Cold War.[4] Trying to predict and shape the impact of future technological innovations is one particularly promising way to improve humanity's long-term future.

This chapter will focus on the long-term impact of artificial intelligence (AI) and the systems that employ it. Many single-purpose technologies, like high-yield crops or traditional weapons, have a general tendency

towards either being beneficial or generating risks respectively. I will argue that because AI can be applied to numerous problems,[5] it has the potential to have an extremely positive impact on the future while also posing risks, with especially substantial risks if AI systems are not developed and deployed with care.

What follows is an introduction to AI and AI policy, and the positive impact that AI safety work can have on the long-term future.

Section 1 provides some background on artificial intelligence and highlights that future AI systems have the potential to have very positive or very negative effects on humanity. Section 2 argues that even though the future trajectory of AI development is uncertain, we can take actions today that can meaningfully affect its trajectory for the better. Sections 3 and 4 present two examples of interventions that can be undertaken now: gathering information about AI progress and impacts (Section 3), and investing resources into the safe development and deployment of AI (Section 4). Section 5 argues that, to do this, we need to avoid situations in which companies or countries are incentivised to invest insufficiently into each of these efforts.

## 1. The past and future of artificial intelligence

There are many technologies with the potential to have a very large impact on humanity. Some of these, such as nuclear weapons and power plants, already exist. Others, such as Dyson spheres, exist only in the minds of science fiction writers. Artificial intelligence systems lie somewhere in between. During the 1970s and 1980s, most AI systems in development were instances of symbolic AI, which used explicit symbols and the manipulation of these symbols to solve complex problems. They applied deductive (that is, mathematical) rules to large bodies of information. People were highly optimistic about what symbolic AI could eventually achieve. This optimism may have been due in part to the fact that symbolic systems are

often good at things that humans find difficult. They can perform complex mathematics and logical inferences with relative ease, for example.[6]

*'Morevac's paradox'* describes the observation that the kind of explicit, abstract reasoning tasks that are relatively difficult for humans to do – such as arithmetic operations involving very large numbers – are often relatively easy for computers. But things humans find relatively easy to do – such as identifying stop signs or playing computer games – are extremely difficult to replicate in a purely symbolic system.[7] Although symbolic AI systems excelled at the former, they never mastered the latter and ultimately failed to live up to people's early expectations.[8]

As funding in symbolic AI systems decreased in the late 1980s and early 1990s, approaches based on artificial neural networks began to increase.[9] Such neural networks are composed of artificial 'neurons' that take multiple values as an input and produce a single output that can serve as the input of one or more subsequent neurons in the network.[10]

Rather than executing a program as a series of manually-specified steps, neural networks are trained to do a task. They are presented with a problem, such as identifying objects in images, and have to learn what patterns to look for in order to correctly identify these objects.

Training methods include supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the model is given labelled data such as images, and must find the best function for predicting future data. In unsupervised learning, the model must identify similarities in data without labels. In reinforcement learning, the model can take actions that are associated with positive or negative feedback.

Compared with the earlier symbolic AI systems, neural networks are much better at the kind of intuitive tasks that humans excel at. This includes tasks like image recognition, playing games, fraud detection, translation, recommendations generation, and language generation.[11]

These systems are improving regularly not only in their capabilities

at specific, narrow tasks, but also in their abilities to learn a wider range of tasks. DeepMind's *AlphaGo* achieved top human-level performance in the game Go, and its successor *AlphaZero* was able to be quickly trained to play not only Go but also chess and shogi.[12] OpenAI's GPT-3 language model was able to achieve state of the art or near state of the art on many natural language processing benchmarks[13] without being fine-tuned to a particular task.

The main factors in improving AI performance are computational power, algorithmic improvements, and more or better data.[14] There are few reassurances about how long the trend of improved performance will continue or how sustainable improvements in these domains are. For example, some believe that increasing fabrication costs will cause a computational bottleneck in the next few years.[15] But we may want to plan for a future in which improvements in AI systems continue, even if we're not certain about the rate of improvement.

*The future of artificial intelligence*

Why do people think that improvements in AI are so important? Crucially, unlike other technologies, AI may in principle be able to do whatever intellectual work humans currently do.[16] AI systems may be able to engineer buildings, prove theorems, provide education, analyse legal documents, write novels, discover new drugs, improve computer security, or produce new music. And they may eventually do each of these things more cheaply than the cost of human labour.[17] Given sufficient information, a sufficiently general system may even have the ability to learn many or most of these tasks, despite not being trained on them specifically.

Future AI systems could be powerful in many respects. First, they could allow many existing tasks to become automated or partially automated, including knowledge work like legal research, medical diagnosis, and decision assistance.[18] Second, they could allow for rapid improvements

in scientific and engineering domains, such as the development of life-extending drugs or building design and materials. Third, they may be able to perform strategically important tasks like cyber-defence, military planning, and election misinformation at a superhuman level. We may come to entrust AI systems with a broader range of tasks and with more consequential tasks over time.[19]

The potential upsides of these systems are large. AI systems could help people lead richer lives. AI systems could make high-quality education available to most people in the form of AI tutors, they could ease day-to-day administrative tasks in the form of AI assistants, they could aid in the development of pharmaceuticals and medical devices that extend and improve life, they could generate new and interesting forms of art, and they could speed up general economic growth in a way that would allow us to distribute more wealth to the global poor.[20]

These systems also come with risks, however. They could be used by bad actors to engage in large-scale misinformation or surveillance, to compromise critical systems, or to create realistic material for blackmail.[21] They could pursue their goals in adverse ways that their users do not intend (as developed in Section 4). They could cause alterations to the economy and to how information is produced that could lead to structural harms, for example from a loss of trust in reputable institutions or the entrenchment of existing social injustices.[22]

## 2. Improving the outcomes of AI under uncertainty

We have already seen promising but limited applications from AI systems; but the AI systems with the largest impact on humanity are likely to be the more capable systems developed in the coming decades. Given our uncertainty about the trajectory that AI research will take, can we meaningfully work on improving the outcomes around the development of future AI systems?

Whether we should try to improve the prospects of a given technology doesn't depend on how far away it is from being developed, but on how much leverage we can expect to have over its impact with the actions currently available to us.

If a powerful technology exists today but we are sure we are unable to influence its trajectory in any way, trying to do so is clearly futile. If a powerful technology will exist far into the future but we are able to influence its trajectory with our actions today, doing so could be extremely valuable.

Typically, the more distant a given technological innovation is, the less valuable it is to try to improve its prospects. This is because the more uncertain we are of when and how something will be developed, the less leverage we have over its trajectory.[23] Precise work to ensure that lunar colonies are built safely is unlikely to be of much value now because we don't have a good enough idea of when and how such things will be built in order to meaningfully assist with safety work now. We also don't know what the background political and economic circumstances will be nor how our current actions will affect them. Moreover, we are likely to have other and greater priorities in the time before the development of this technology.

It would be a mistake, however, to throw up our hands and assume that just because the trajectory of some technology is long and uncertain, there is nothing we can do to influence it. When we are making decisions under deep uncertainty, we can do things like:

1. Gather information about high-stakes outcomes: take actions to identify broad clusters of possible futures, their value, and our degree of influence over them.

2. Gather information about the technology: for example, about the technology itself and its impact.[24]

3. Identify and perform robustly good actions: take actions that improve

the situation in the highest-stakes outcomes, that leave more options open to us in the future, and that result in outcomes that are broadly good even if the outcomes of these actions are unlikely to be optimal.

This kind of decision-making under deep uncertainty has been explored more formally in the context of 'robust decision-making'.[25]

Important developments in AI do not seem so far off or unpredictable that we cannot have any real influence on them. But they are also not so predictable that we can afford to ignore the heuristics of robust decision-making. In the next three sections I will give some examples of actions that we can take to improve the prospects of future AI development.

### 3. Tracking the progress and impact of AI

When we are operating under uncertainty, the value of information increases. The more information we can gather, the more informed our future decisions will be. One of the key difficulties for anyone who wants to develop good policy for AI future systems is that we don't have a clear enough picture of AI development and its future impact on society.

In order to get information about how future AI systems are likely to impact society, we need to develop infrastructure that can measure the impact of current AI systems on society and that can predict the impact of AI systems currently being developed. This involves tracking (i) improvements in existing systems across a range of tasks, (ii) what components are contributing to these improvements and to what extent, (iii) how safe and secure existing systems are relative to their capabilities, and (iv) how improved AI systems are impacting society via the economy, political institutions, and more.

We can roughly divide these uncertainties into those that are about the development of AI systems and those about the impact of AI systems.

The first source of uncertainty concerns the rate and nature of the development of AI systems themselves. There are a variety of machine

learning benchmarks that measure capabilities at specific tasks such as image recognition and natural language inference.[26] It will become increasingly important to construct benchmarks that can measure more economically important tasks that require a broader range of skills to complete, and to construct benchmarks that can better measure the breadth and robustness of AI capabilities.

The development of standardised AI benchmarks that can capture the success of systems at economically and strategically important tasks is something that could be done by academics in collaboration with governments, non-profits, and industry.

The second source of uncertainty concerns the impact of AI systems. To give one example, there is currently much uncertainty about the impact that the incremental or sudden automation of important tasks could have on the economy and on society as a whole.[27] How might the rapid automation of truck driving or paralegal work affect society?

We can gather information on the current and potential impact of automation and the best interventions for preparing a workforce that might need to adapt to these shifts. Governments and policymakers may want to improve their ability to track and respond to potentially rapid automation across industries.[28] This is especially important because it is not guaranteed that automation will occur in an incremental fashion within specific industries: once key tasks have been mastered by AI, it may be possible to automate tasks across multiple occupations or industries at once.[29]

Another example of our uncertainty about the impact of AI systems is how often they are being misused and how often accidents are occurring. Governments, non-profits, and industry actors can take steps to monitor accidental harms from AI, such as deaths resulting from automated vehicles, major misdiagnoses by ML-assisted diagnostic tools, or failures in safety-critical AI systems. They could also take steps to predict and

track the malicious use of AI – though these may be harder to track in real time as bad actors attempt to avoid detection – and structural harms like labour displacement, rising inequity, or the erosion of privacy.[30]

Given this uncertainty, it will become increasingly important to measure the safety and security of AI systems as they play a more important role in society. This will be critical in the development of good standards for deployable AI systems. For example, NIST research currently focuses on how to measure the security and explainability of AI as part of its efforts to develop AI standards.[31]

Although work is increasingly being done to measure both the impact and the development of AI, there is still much more that can be done to decrease our uncertainty in these domains.

**4. Work on the safe development and responsible deployment of AI**

*The safe development of AI*

AI systems are 'aligned' with human values if their goals are in accordance with the intentions of the user and with the broader constraints set out by society as a whole, such as acting within the law.[32] They are 'safe' if they are unlikely to cause accidental harm and are not easy to misuse.

Part of the reason that modern AI systems are not safe and aligned by default is that they are trained to carry out hard-to-specify tasks rather than executing simple algorithms. The method that the systems are using to execute these tasks after training is often not easy to identify even if one has full access to the model, and it is therefore difficult to predict whether it will behave as desired all the time or in unusual circumstances.[33]

For example, we may want a language model to avoid offensive language when writing children's stories. But although we can take steps to reduce the likelihood of this, it is hard to eliminate the possibility entirely. Suppose we can find a way to remove specific offensive words from its

outputs. A lot of very offensive language does not require the use of such words. The phrase *'unattractive people should not be allowed to have children'* doesn't contain any particularly offensive words, but it is not the kind of standalone phrase we would want to see in a children's story. Since we can't predict what a language model's training will cause it to output in response to a given prompt, it is hard to completely eliminate the possibility of this phrase appearing even in response to prompts not designed to elicit biased outputs.[34]

The more complex a task a system can do, the harder it is to predict precisely how it will carry out the task. The harder it is to predict precisely how a system will carry out a task, the more guarantees we need that it will not act in ways that are radically unexpected and harmful.

Ensuring that AI systems are safe and aligned increases in importance the more powerful AI systems become. It may also increase in difficulty: it could be especially difficult to ensure that systems are safe if they are more capable than humans across a range of domains.

AI safety is a field that is trying to make safe and aligned AI by reducing the likelihood of things like unsafe exploration and 'reward hacking'[35] – a system's pursuit of specified goals through undesired means – and finding ways to align AI systems with what humans want.[36] It includes specific proposals for this, such as 'AI safety via debate',[37] in which ML agents learn human reasoning and values by adversarially training to debate various questions. Since humans judge the outcomes of these debates, the ML agents must learn what reasoning and considerations the humans find compelling in order to win.[38]

We can strengthen the AI safety discipline by funding independent safety research at non-profits and academic institutions.[39] Governments can also incentivise industry actors to invest more in safety work by creating international guidelines for safe AI development or rewarding industry actors that invest heavily in safety.

*The responsible deployment of AI*

In addition to safe development, we must deploy AI responsibly. Responsible deployment of AI involves at least four things.

First, it involves developing ways of testing whether systems are sufficiently safe and aligned and releasing those systems with due care, for example by incrementally releasing models in order to get information about accidents, misuse, or systemic harms that might result from them[40] or releasing model cards alongside the models that describe possible limitations, biases and misuses.[41] Second, it involves taking steps to ensure that the model cannot be stolen or misused by bad actors. Third, it involves contributing to a safety-oriented development environment that doesn't incentivise a race to the bottom on safety (as discussed in the next section). And fourth, it involves ensuring that the economic benefits generated by AI systems are distributed universally.[42]

Responsible deployment of AI systems may be able to help us tolerate some deficiencies or setbacks in the work on safe development by ensuring that no one deploys systems until the safety research has caught up with their capabilities.

Safe development of AI systems may help us to tolerate some deficiencies or setbacks in responsible deployment, by ensuring that the systems being deployed are not likely to result in accidents, even if there is not yet a good governance structure in place that can prevent such systems from being deployed at all.

Industry, governments, non-profits, and academia can all contribute to the safe development and deployment of AI systems. For example, governments can fund research and support new solutions to ensure safe AI deployment, such as global regulatory markets for AI safety.[43]

The best safety solutions will be rooted in a deep understanding of the economic and political threats to safe and aligned AI. This is the subject of the next and final section of this chapter.

**5. Preventing collective action problems**

One possible threat to the safe development and responsible deployment of AI are collective action problems. Collective action problems are situations in which the collective would be better off if everyone cooperated with one another, but it is in each individual's interest not to cooperate. Many of the challenges facing those who want to develop AI safely can be characterised as collective action problems.

First, let's consider the collective action problems that companies face. Work on the safe development and responsible deployment of more capable AI systems is costly in time and money. In most industries, companies are incentivised to invest in safety by market incentives, liability laws, and regulation.

These three types of incentives also apply to companies developing AI systems, but they appear to be weaker here than in other industries. Consumers are less likely to be able to evaluate the safety of AI products initially, which weakens market incentives to develop them. It is unclear how current liability laws will apply to AI systems, and smaller companies may not have the funds required to pay for harms they are held liable for. Regulators often don't have the kind of knowledge of AI systems that is required to construct effective regulation, and may find it difficult to keep up with the pace of change in AI development.[44]

Developing safe products is more valuable to companies than developing unsafe products, all else being equal. But developing safe products requires an investment of time and money that companies might not want to incur if their competitors do not. After all, it's often beneficial for companies to release their products sooner and be first to market.[45] So it can be in the interest of each individual company to defect: to underinvest in safety in order to put more resources into releasing their product sooner. This can be true even if each developer would much prefer to compete in an environment in which developers must all invest more resources in safety.

In this kind of "race to the bottom" on safety, raising the bar on what companies must invest in safety can result in a situation where each company is strictly better off. By raising the bar for everyone, each company would expect to create a more valuable end product without adopting a competitive disadvantage relative to the low safety investment scenario. But how do we raise the bar on what companies must invest in safety? This is the collective action problem for AI safety in industry.[46]

This problem is compounded by the fact that similar collective action problems exist between governments and their regulatory bodies. Governments have an incentive to support the development of more capable AI systems by companies within their borders, since these systems likely drive economic benefits. But the AI systems that are developed by a company in one country can have wide-ranging effects in other countries that are difficult for them to prevent. For example, a system that is designed in one country can be used to create convincing political misinformation that is primarily targeted at the citizens of another country.

It is therefore in the interest of all governments and regulators to agree to a minimum set of standards for the safe development and responsible deployment of AI systems, since this reduces the likelihood that accidents or misuse involving AI will negatively affect them, without reducing their relative advantage when it comes to developing their own AI capabilities. This requires international cooperation on the safe development of AI.

There are steps we can take to increase the likelihood of cooperation on the safe development of AI internationally and across industry. These include identifying opportunities to collaborate on shared research projects,[47] increasing the transparency into AI development, and increasing levels of trust that others are developing AI safely.[48]

Identifying and dismantling collective action problems that get in the way of safe AI development is one potentially robust way to improve the trajectory of AI.

## 6. Conclusion

At the beginning of this chapter I noted that technological innovation has been a cause of both great flourishing and great risk in the history of humanity. I then argued that what could set artificial intelligence apart from other forms of current technology is that it may be able to perform all kinds of intellectual work. If this promise is realised then the potential upsides of AI for humanity in the long term could be extremely high, but so could its long-term potential to be misused or to create serious unintentional harm if it is not developed safely. There is still much uncertainty about both the trajectory and the timeline of AI development. This can make work to improve the prospects of AI development seem intractable. However, there can still be robustly good steps available to us even in situations of great uncertainty. Work to track the progress and impact of AI, progress on the problem of safely developing and responsibly deploying AI, and actions that reduce collective action problems around the safe development of AI all seem like plausible candidates for robustly good steps in AI development.

My goal in this chapter has not been to offer specific advice for those interested in working on or investing in these areas: resources for this exist elsewhere.[49] Instead, I have tried to show that ensuring the safety of artificial intelligence is something that could have a significant long-term impact on humanity, and that there is a case to be made for beginning this work now.

# IV

# Towards a Longtermist Culture

CHAPTER ELEVEN

# Traversing the Garden of Forking Paths More Wisely: The Challenges of Taking the Long Term Into Account in Decision-Making

by
Hiski Haukkala
*Professor of International Relations, Tampere University, Finland*
*and Associate Fellow, Europe Programme, Chatham House*

*This chapter discusses the challenges of inserting long-term thinking into the way we deal with current affairs and decision-making. The fact that our decision-making remains mired in short-term considerations is not an accident. Human individuals and societies have built-in constraints — biological, psychological, political and institutional — that make operating in a fully rational and far-sighted manner very difficult. But taking the long term into account is not only preferable, it is imperative due to increasingly mounting catastrophic and existential risks. This chapter considers various political reforms as solutions to short-termism, and argues that while such reforms are imperative, they are also insufficient. What is ultimately needed is a deep cultural change that empowers and even compels us to factor the long term into the very fabric of our lives. To this end, a new cultural tradition is proposed for the whole of humanity, one which reminds us to be custodians of this planet and its long-term potential so that we can together ensure a viable, open, aspirational future.*

Why do so many burning issues of the day seem to catch us unawares and turn into rapidly escalating crises?[1] Are there ways to improve our decision-making processes, and to increase our chances of dealing with problems preemptively to avoid the most negative outcomes that might threaten the very future of humanity? How could we spur change in the right direction?

This short chapter seeks to address these questions by discussing the challenges of inserting long-term thinking into the way we deal

with current affairs and make decisions. The challenges stem from many different sources and achieving improvement will be far from straightforward: The fact that our decision-making processes remain short-sighted and mired in short-term considerations is not an accident. Human individuals and societies are faced with several built-in constraints – biological, psychological, political and even institutional – that make it very difficult for us to operate in a fully rational and far sighted manner.

It must be said, however, that our standard operating procedures have served us very well. We have harnessed the forces of nature and built systems of bewildering complexity. In the process, almost the whole of humanity has grown more prosperous and secure. Life expectancy has soared, and individuals have been empowered almost throughout the globe. Our collective progress is hard to deny, and to a degree this success has become a factor that blinds us to considering alternatives, even as evidence mounts that the current course of action is not sustainable.

Indeed, our success has been built on a way of life that is now eroding other pillars of our well-being and potentially even our species' existence. Climate change is advancing rapidly, and the biosphere is under alarming strain. Exciting technological advances are bringing significant gains and potential solutions to many problems, but we know from history that such progress also always results in new and often entirely unanticipated problems and sometimes great dangers.[2] At the same time the very complexity and interconnectedness of our economic, social, political and technological systems creates new risks.[3] Consequently, the probabilities of global catastrophic risks are mounting.[4] Our propensity to focus on the issues of here and now is jeopardising the quality and perhaps the very existence of our long-term future.

Our current approach to decision-making is part of the problem. It seems clear that it is ill-suited to dealing with the mounting challenges. The action-reaction cycle that we see today is conducive to international

conflict and results in suboptimal responses to other challenges and problems. At the same time, our current approach acts as an incubator of sorts for systemic risks; instead of solving them, it generates and prolongs them while making us blind to their very existence.

To get out of the present rut we need more long-term oriented thinking, analysis and planning. The road to better decisions based on more appropriate time frames will not, however, be easy. The limitations in our biology, psychology, politics and institutions practically ensure the continuation of short-termist and suboptimal policies and responses in our current affairs. We should, however, bear in mind that there are also legitimate forms of short-termism.[5] We exist in, and can therefore only act in, the moment in time in which we find ourselves. The challenge is to devise forms of politics and institutions that embed long-term considerations into our everyday comings and goings.

To chart an alternative way forward, this chapter seeks to expand the notion of 'current affairs'. As the old saying goes, 'the urgent crowds out the important' in politics. We must appreciate that current and urgent issues may often not be the most important after all, and at times we must prioritise long-term issues even if they may not seem pressing in the moment.

Those who have written on the topic of short-termism have often proposed reforming our institutions to address the problem.[6] Such reforms are useful and indeed necessary, but for them to work as intended we need a mindset change, and that can only be achieved by a fundamental cultural shift, one that I will argue for in the conclusion to this chapter.

But before proceeding any further, a brief note concerning long-termism is in order. In business, a strategy that spans one to ten years is often called a long-term strategy. In politics, as will be argued later, the time frame can be even shorter. This chapter departs from an understanding of long-term futures as having a lower bound of a generation – approximately twenty-five years. On many issues, a longer time frame is called for. In

terms of our current politics and decision-making, a shift to this scale would be a revolutionary change.

## 1. The Garden of Forking Paths

In 1941, the Argentinian writer Jorge Luis Borges published a short story called The Garden of Forking Paths. It involves the idea that instead of choosing between alternative futures that preclude all others at the point of taking a decision, humans in fact create an infinite number of different forking realities in their lives. His story inspired multiverse theories, and it can also serve as a point of departure for this chapter. Following Borges; for humans, our futures, even the faraway ones, come to us one event and decision-making point at a time. We experience our lives as an endless chain of moments and decision points, seamlessly flowing one after another.

Our way of experiencing the flow of time as 'one damn thing after another' is natural: as was already noted, we mainly exist in the moment that we can experience through our senses and can only physically interact with the world as it is here and now. This seems obvious, even banal, but it yields two important insights that are relevant when thinking about our current decision-making.[7]

Firstly, on the positive and empowering side, this observation opens pathways in the form of long causal chains for effecting long-term change. Human agency matters, and if exercised with foresight and wisdom it can be used to foster desirable long-term futures – and to avoid undesirable ones. Through our imagination we can gain a kind of access to the future and envision alternative realities, assess their likelihood and desirability, and act accordingly.8 We can also conduct research and analysis that might give us a better understanding of current trends and their long-term effects and trajectories. We can extrapolate current trends but should keep in mind that not every issue of significance grows linearly or is easily foreseen.

Secondly, however, this observation reveals the daunting complexity of choice. The future is generated one step and choice at a time. One need not subscribe to the multiverse theory to appreciate how over time this results in decision trees of bewildering intricacy even in our personal lives, let alone on a global level. Yet we need not, and indeed must not, remain passive passengers on our way to our rendezvous with the future. We should prune this garden of forking paths to our advantage and chart a course that is as beneficial as possible, while seeking to avoid outcomes that could turn out to be catastrophic.

The call to optimise our global decision tree opens two problematic elements, however. When thinking about the paths pursued, we must always ask to whom they are beneficial. Issues of debate can very rarely be tackled purely analytically. Most of these issues are also inherently political and thus involve – and to a degree also revolve around – the interests and power of key actors. Indeed, the vested interests at stake often skew the process. The issue is never that of policy optimisation, but rather that the eventual policies and the framing of the issues are an outcome of a political process that is always in the final analysis a struggle for power and for the eventual nature of solutions chosen. Any choices made will generate both winners and losers and it is erroneous to view them as questions of policy optimisation void of politics and struggle for power. In fact, it is hard to envisage a long-term future that would be equally beneficial to everyone. The distribution of relative gains and the competition concerning who gets to call the shots is the bread and butter of politics. This applies at both the domestic and global level and the short and the long time frame.

The second problem is that this naturalised flow of time may blind us to the possibilities of forging a path to desired long-term futures. This is partly because we simply lack a sufficiently sophisticated understanding of our world and the ability to try and effect desired change with a high enough degree of certainty. But perhaps more importantly, we are

often blinded by current events, unable to see beyond the situation at hand. Related is the so-called *'normality bias'*, the idea that we anchor our expectations of the future so firmly in the *status quo* that we find it difficult to fully comprehend possible catastrophe.[9] As will be argued below, this is increasingly a threat to the prosperity and even the very survival of humanity.

Fostering decision-making that considers long-term challenges and effects would be rational and desirable. There is evidence that individuals who postpone immediate consumption in favour of long-term investments fare better than those who operate on a more short-term and hedonistic outlook.[10] It seems reasonable to expect the same at the level of society. But taking the longer time frame into account is not only preferable, it is imperative. Toby Ord has argued that catastrophic and even existential risks from causes such as unaligned AI, climate change tail events, nuclear war, and several others, are rising.[11] In his assessment, the odds of a disastrous outcome for humanity – defined as the extinction of the species or a drastic and terminal curbing of its potential – within this century could be as high as one in six. This figure obviously comes with a very high degree of uncertainty but even a significantly lower risk of catastrophe would be unacceptable.

It is simply irrational to continue ignoring these risks. Yet the bad news is that policymaking is never fully rational. The idealised version of decision-making prevalent in many rational choice theories does not in fact capture the messiness of real human decision-making processes (this was of course never the point of these theories to begin with). Herbert Simon famously coined the term *'satisficing'* to capture the fact that most of the time people settle for decisions and outcomes that are *'good enough.'*[12] But what might seem good enough in the short-term can prove disastrous in the long-term. The law of unforeseen and unintended consequences often prevails.

**2. The biological constraints and political and institutional dilemmas**

Causes of irrational long-term decision-making abound, and find their roots in many features of human life. One such cause stems from the hectic pace of everyday politics. Even in normal times a player in this game is often rushing after the ever-present now. The essence of this dynamic has already been captured by Whiting, writing five decades ago (1972, 236): for a bureaucrat, '*tomorrow's problem can be taken up next week but today must still be devoted to yesterday's agenda. Unfortunately, the today is always with us; next week never seems to come.*'[13] My own experiences working for the Ministry for Foreign Affairs of Finland seem to verify this observation. Although I was initially hired to conduct long-term strategic thinking and planning, my presence was soon seen as a spare resource to be used in tending to other more immediate items on the ministry's agenda. I do not complain, as the experience of being closely attached to the policy process was very rewarding, but the experience seems to suggest that diplomatic services do not really know how to use slower forms of long-term thinking in their own, much faster processes. Foreign policy is not all crisis decision-making, yet the civil servants are chronically short of time: they simply have too much on their plate.[14]

Increasingly all government departments all over the world are hindered by the same problem. The standard operating procedures are short-termist by nature and prone to deviating from the ideal of rationality.[15] Our increasingly crisis-prone world compounds the problem, inviting and forcing us to shrink our time horizons at a time when it would be more important than ever to expand them. We are stuck in a reactive crisis-management mode where we are even less able to see the bigger picture and think of the long-term challenges and the consequences of our actions. One needs only to look at our responses to the COVID-19 pandemic and its effects on our national and international politics to see the point.

But the seemingly pressing nature of our current affairs is not the only factor ensuring suboptimally short time frames. As a species, Homo sapiens is practically hardwired not to be able to think about issues in the long term. Our ancestors operated in an environment that mainly consisted of issues of immediate benefit or danger. As a consequence, as a species, we have evolved to think and act accordingly. Indeed, we often tend to forget that humans are only one mammal among many in the world. We have evolved culturally very rapidly but biologically we have remained largely intact for the last 40,000 to 50,000 years. It is not an exaggeration to say that biologically we are an incredibly large and unwieldy host of hunter-gatherers that have been thrust into the bewildering complexity of the 21st century.

Luckily, biology is not the only thing that bears on evolution. Ever since the advent of writing, we have been able to transcend the cognitive limits of single individuals or even generations, passing on information over vast distances and long time periods. This has enabled a rapid rise in science and technology and the many benefits they bring – benefits we often take for granted. Through this cultural evolution we have been able to develop increasingly complex systems of national and international governance. But at the same time, those systems have not been able to reliably produce just and effective solutions to humanity's greatest problems.

On the contrary, our politics and political decision-making remain short-sighted and often mired in zero-sum games. To begin with, the basic organisational unit in the world is still a sovereign nation state, although the issues we face increasingly know and respect no boundaries. Moreover, our propensity to think in terms of national welfare and security is increasingly at odds with the common fate of humanity. The GDP might be a good metric for keeping tabs on economic output within a certain segment of the world, but it is clearly insufficient as a guiding principle

for our collective responses. To be clear, I am not advocating abolishing nation states, as they remain indispensable reservoirs of political authority, competence and legitimacy, but an outlook more attuned to our common challenges is clearly called for.

A related problem is that the time horizons of governments often stem from their expectations concerning how long they expect to be able to govern and how their policies might affect their chances of getting reelected or, in the case of authoritarian regimes, keeping hold of power.[16] This means that, though usually short, time horizons are not uniform across governments. To give two extreme examples, the Communist Party of China assumes that it will be in power indefinitely – and to a degree it operates accordingly: it currently has development plans for the country that span up to the year 2049, the centenary of the party. By contrast, at the time of writing (summer 2020) the US President Donald Trump is fighting for his political survival and does not show any propensity for thinking beyond the tweet at hand.

It is hard to generalise on these time frames, at least without more systematic research on the topic. Yet it seems safe to conclude that most governments are thinking in time frames of months and years, not decades, let alone generations or millennia. Even in the Chinese case it can be argued that despite their rhetoric of decades-distant goals, their actual decision-making often reflects the needs of the moment – chiefly their short-term economic ascent.

### 3. Conclusions: The need for a cultural change

A change to the way we make decisions is clearly called for. The challenges and risks are rising and the need for this change is increasingly urgent. Yet the analysis here points to a grim conclusion: the inadequacy of our current policies and decision-making seems to be hardwired into our biology and institutions. The escalating crises are inviting us to reconsider

the ways we run our current affairs and to better prepare for the future, but there is no guarantee that humanity will embrace the right change in time to avert catastrophe.

Does this mean that we are doomed? The call to orient policy more to the long-term is old hat. If this would come easily or naturally to us, surely the problem that a lot of people already recognise would be solved by now?

This indicates that there are structural factors impeding the change. But what humans have created, they can also amend. Parliamentary committees for the future are often praised as potential solutions. My own country, Finland, is known for having a Committee for the Future at its parliament. The role of the committee should not be dismissed.[17] It is imperative that these perspectives are brought to bear on our political debates. At the same time, the parliament, and perhaps the committee in particular, is simply too far removed from the day-to-day policymaking to make a crucial difference.[18] We also know that this is a typical problem with governmental foresight processes, and in any case they often do not hold a long enough perspective. Although the objectives behind these efforts are commendable, the outcomes leave a great deal to be desired.

The persistence of these difficulties has led some people to become exasperated with the lack of foresight and wisdom in our politics and to start flirting with the idea of authoritarianism as a potential solution. Yet there is very little evidence that authoritarian governments are more interested in the far futures or even in the well-being of their own people than democracies.[19] Indeed, here I echo the words of Winston Churchill, who famously argued how 'it has been said that democracy is the worst form of Government except for all those other forms that have been tried from time to time'. The sensible way forward will not be a turn towards authoritarianism, however enlightened a form of authoritarianism it might purport to be, but making our democratically elected leaders more

attuned to the long-term consequences of their decisions.

This turn has already been advocated in the case of leadership.[20] Addressing the leaders is a good start but it will not be enough on its own if the incentive structures of our current politics remain intact. In democratic systems especially, leaders will not be empowered to take the long-term point of view more into account without popular support. And gaining that support entails an electorate that would both demand and reward responsible long-term thinking and policies and punish the lack of them.

This is easier said than done and in order to get there we need cultural change. We must all learn to take the long-term much more seriously and factor it in at the very fabric of our lives. This fundamental mindset change could only begin to be achieved by alerting people to their responsibilities. The so-called 10,000 Year Clock – an art installation, currently in production, consisting of a clock designed to tick for 10,000 years – is one such prompt. The aim is commendable, though the clock remains at least in its present guise more a gimmick and perhaps even a distraction than a solution to the problem at hand.[21]

A better solution might be to adapt the Canadian tradition of 'The Ritual of the Calling of an Engineer' in which engineers are bestowed with iron rings to remind them of the heavy responsibility their profession entails. We should consider adapting this tradition to the whole of humanity. Let us call it 'The Stone Ring of Longtermists' and it should be given to every individual at birth to signal that we, the currently living, are not here only at and for our own leisure but that we are also the custodians of this planet and its long-term potential and that each and every one of us should feel that responsibility and act accordingly.[22]

To that end, three words should be inscribed on the ring: 'viable', 'open', and 'aspirational'. They would refer to three basic rules of thumb that should be observed by humanity going forward:

1. Keep the long term viable. Deal with the current threats we face and know of (and even those we are not aware of) so that we can avoid catastrophe for humanity and this planet.

2. Keep the long term open. Avoid choices and decisions that would constrain our freedom in the future. In particular, we should retain the ability to undo and *'unchoose'* things and we should therefore avoid undesirable lock-in technologically, economically or politically.[23] Obviously, we can never fully walk back developments and decisions taken but we should retain the ability to undo the most harmful or negative aspects of said choices, should we so desire. In addition, we cannot infer with any certainty the preferences of future generations apart from the fact that most probably they would like to exist and get to realise their own potential. This, too, underlines the importance of the first two rules of thumb.

3. Finally, keep the long term aspirational. We need long-term thinking not only to avoid bad outcomes or even outright catastrophes but to envisage and realise positive and just futures for us all, including future generations.

Most importantly, we must embrace the long term in all aspects of life. It is only by doing so that we can successfully traverse the garden of forking paths and not fall drastically short of our potential.

# END NOTES

### Introduction

1. Thanks to Toby Ord for this fact, via Will MacAskill.
2. Indeed, the Earth's total historical biomass is sixteen times lighter than the amount of gold produced from a single collision between two neutron stars. (Assuming the current annual production rate of biomass of 104.9 billion tonnes per year for all 3.5 billion years of life on earth, and six billion trillion tonnes of gold produced from one such collision.) C. Cookson, Scientists Discover the Origins of Gold in Space, *Financial Times,* 2018; C.B. Field, M.J. Behrenfeld, J.T. Randerson, P. Falkowski, Primary production of the biosphere: integrating terrestrial and oceanic components, *Science* 281 (1998) 237-40.

### 1. Expanding the Moral Circle to Future Generations

1. Thanks to Tyler John for invaluable contributions to this article.
2. H. Cunningham, The multi-layered history of Western philanthropy, *The Routledge Companion to Philanthropy*, ed. T. Jung, S.D. Phillips and J. Harrow (2016).
3. P. Vallely, *Philanthropy from Aristotle to Zuckerberg,* 2020, 290.
4. Vallely, *Philanthropy A to Z*; James Baldwin Brown, *Memoirs of the Public and Private Life of John Howard, the Philanthropist*, London, 1823 (2nd edn), 19-21.
5. G.F.R. Barker, Howard, John (1726?-1790), entry in *Dictionary of National Biography* 28 (1885-1900).
6. Vallely, *Philanthropy A to Z*, 291.
7. Vallely, *Philanthropy A to Z*, 292.
8. J. Field, *The Life of John Howard With Comments on His Character and Philanthropic Labours,* 1850, 344.
9. Vallely, *Philanthropy A to Z*, 298.
10. 'Elizabeth Fry', in the *Howard Journal of Criminal Justice* 16.
11. The 1823 Act further eroded prisons' perverse profit incentives by providing payment for jailers, required female and male prisoners to be separated and placed female wardens over female prisoners, prohibited the use of irons and manacles, and lifted the death penalty from 130 crimes. The 1835 Act introduced paid prison inspectors and provided prisons with financial assistance from the Treasury, to ensure the enforcement of the 1823 Act.
12. A.Y. Davis and C. Shaylor, Race, gender, and the prison industrial complex, *Meridians* 2.1 (2001) 1-25.
13. C. Long and K. Fingerhut, AP-NORC poll: Nearly all in US back criminal justice reform, *AP,* 2020, https://apnews.com/article/police-us-news-ap-top-news-politics-kevin-richardson-ffaa4bc564afcf4a90b02f455d8fdf03
14. We derive the idea of a "moral circle" from Peter Singer, *The Expanding Circle*, 1981.
15. 15.A.Nellis, The colour of justice; racial and ethnic disparity in state prisons, *The Sentencing Project*, 2016, https://www.sentencingproject.org/publications/color-of-justice-racial-and-ethnic-disparity-in-state-prisons/
16. M. Schreiber, What kills 5 million people a year? It's not just disease, *NPR,* 2018, https://www.npr.org/sections/goatsandsoda/2018/09/05/644928153/what-kills-5-million-people-a-year-its-not-just-disease
17. K. Anthis and J. Reese Anthis, Global farmed and factory farmed animal estimates, *Sentience Institute,* 2019, https://www.sentienceinstitute.org/global-animal-farming-estimates
18. M. Roser, Number of democracies, *Our World in Data*, 2019, https://ourworldindata.org/democracy#number-of-democracies
19. Labor force participation rate, female (% of female population ages 15+) (modeled ILO estimate), world, *The World Bank*, 2020, Databank.https://data.worldbank.org/indicator/sl.tlf.cact.fe.zs
20. F. Ting, Z. He, and R. Baillargeon, Toddlers and infants expect individuals to refrain from helping an

ingroup victim's aggressor, *Proceedings of the National Academy of Sciences* 116 (2019) 6025-6034.
21. J.K. Hamlin, N. Mahajan, Z. Liberman, and K. Wynn,. Not like me = bad: Infants prefer those who harm dissimilar others, *Psychological science* 24 (2013) 589-594.
22. A. Buchanan and R. Powell, *The Evolution of Moral Progress: A Biocultural Theory*, Oxford, 2018; K. Sterelny, *Evolutionary Foundations for a Theory of Moral Progress?*, 2018.
23. Buchanan and Powell, *Not like me*; Sterelny, *Evolutionary Foundations.*
24. Kim Sterelny attributes this plasticity to the need to make norms explicit as societies grew larger and more complicated; making norms explicit made it easier for our ancestors to modify these norms and internalise novel norms.
25. R. B. Lee, Hunter-gatherers on the best-seller list: Steven Pinker and the "Bellicose School's" treatment of forager violence, *Journal of Aggression, Conflict and Peace Research* 6 (2014) 216-228; Sterelny, https://bigthink.com/culture-religion/what-started-poverty?rebelltitem=1#rebelltitem1
26. Singer, *The Expanding Circle*, 112.
27. Estimated number of African slaves who were taken from various regions of Africa during the transatlantic slave trade in each century from 1501 to 1866, *Statistica*, https://www.statista.com/statistics/1150475/number-slaves-taken-from-africa-by-region-century/
28. N. Nunn and L. Wantchekon, The slave trade and the origins of mistrust in Africa, *American Economic Review* 101 (2011) 3221-52; E. Green, Explaining African ethnic diversity, *International Political Science Review, 2012,* https://journals.sagepub.com/doi/full/10.1177/0192512112455075; N. Nunn, Understanding the long-run effects of Africa's slave trades, 2017, https://voxeu.org/article/understanding-long-run-effects-africa-s-slave-trades: There is evidence that Southern U.S. counties with higher rates of slave ownership before abolition have higher rates of economic inequality between races today, more hate crime, higher rates of racist attitudes, lower rates of support for affirmative action, and more; G. Bertocchi, The legacies of slavery in and out of Africa, *IZA Journal of Migration* 5 (2016); C. Gunadi, The legacy of slavery on hate crime in the United States, *Research in Economics* 73 (2019) 339-344; A. Acharya, M. Blackwell and M. Sen, The political legacy of American slavery, *Journal of Politics* 78, (2016) 621-641. Thanks to Will MacAskill for these facts.
29. M. Rediker, *The Fearless Benjamin Lay,* 2017. Thanks to Will MacAskill for these facts.
30. J. Henrich, *The WEIRDest People in the World,* 2020.
31. Alliance for Charitable Reform,http://acreform.org/paw/philanthropic-achievement-of-the-week-american-anti-slavery-society/
32. Charity Aid Foundation, From the margins to the mainstream, 2020, https://www.cafonline.org/docs/default-source/about-us-policy-and-campaigns/from-the-margins-to-the-mainstream--philanthropy-diversity-equity-and-inclusion-in-our-society-june-2020.pdf
33. Charity Aid Foundation, Margins to mainstream.
34. B. Mandeville, An essay on charity and charity-schools*, The Fable of the Bees,* 1732.
35. A. Davis, *Are Prisons Obsolete?,* 2003, 80-81.
36. Anthis and Reese Anthis, Farmed animal estimates.

**2. The Moral Case for Long-term Thinking**

1. R. Nigel, *Brewer's Quotations: A Phrase and Fable Dictionary*, London, 1994.
2. P. E. Tetlock, *Expert Political Judgment*, Princeton, 2005.
3. W. MacAskill, Longtermism, *The Effective Altruism Forum* (blog), 2019, https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism; H. Greaves, and W. MacAskill, The case for strong longtermism, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism/
4. F. Adams, Long-term astrophysical processes, *Global Catastrophic Risks*, Oxford, 2008.
5. F. Adams 2008.
6. H. Greaves, Discounting for public policy: A survey, *Economics & Philosophy* 33, no. 3 (2017) 391-439 surveys further arguments for and against a positive rate of pure time preference.
7. S. Frederick, G. Loewenstein and T. O'Donoghue. 2002, Time discounting and time preference: A critical review, *Journal of Economic Literature* 40, no. 2 (2002) 351-401.
8. Nick Bostrom presents this argument in more detail. See: N. Bostrom, Existential Risk Preven-

tion as Global Priority, *Global Policy* 4, no.1 (2013) 15-31; N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*, Oxford, 2014.
9. P. Millett and A. Snyder-Beattie, Existential risk and cost-effective biosecurity, *Health Security* 15, no. 4 (2017) 373-83.
10. www.centerforhealthsecurity.org/
11. www.fhi.ox.ac.uk
12. This slogan is a rephrasing of J. Narveson, Moral Problems of Population, *The Monist* 57, no. 1 (1973) 62-86; For a more recent discussion of person-affecting approaches see: M. A. Roberts, An asymmetry in the ethics of procreation, *Philosophy Compass* 6, no. 11 (2011) 765-76.
13. Thomas Teruji canvasses a range of possibilities in T. Thomas, The asymmetry, uncertainty, and the long term, 2019, https://globalprioritiesinstitute.org/teruji-thomas-the-asymmetry-uncertainty-and-the-long-term/
14. T. Hurka, Asymmetries in value, *Noûs* 44, no.2 (2010) 199-223.
15. D. Althaus and L. Gloor, Reducing risks of astronomical suffering: A neglected priority, *Center on Long-Term Risk* (blog), 2019, https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/
16. Bostrom 2014, chapters 1-2; V. C. Müller and N. Bostrom, Future progress in artificial intelligence: A survey of expert opinion, *Fundamental Issues of Artificial Intelligence*, 553-571, 2016; K. Grace, J. Salvatier, A. Dafoe, B. Zhang and O. Evans, When will AI exceed human performance? Evidence from AI Experts, *Journal of Artificial Intelligence Research* 62 (2018) 729-54.
17. R. Hanson, *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford, 2016, 57-58.
18. D. J. Chalmers, 2010, The singularity: A philosophical analysis, *Journal of Consciousness Studies* 17 no. 9-10, 7-65; Bostrom 2014; T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, London, 2020.
19. www.openAI.com
20. www.cset.georgetown.edu
21. N. Stern, *The Economics of Climate Change: The Stern Review*, Cambridge, 2007; R. S. Pindyck, Climate change policy: what do the models tell us? *Journal of Economic Literature* 51, no.3 (2013) 860-72.
22. IPCC. 2014. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change,* Cambridge, UK and New York, NY.
23. H. Greaves and W. MacAskill, The case for strong longtermism, 2019, https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/
24. This figure comes from GiveWell's 2020 charity assessments. Their model can be found here: www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models. Note that GiveWell's focus is on health in the developing world. Short-term initiatives aimed at helping animals are plausibly even more cost-effective. See www.animalcharityevaluators.org for more.
25. P. Trammell, Discounting for patient philanthropists, 2020, https://philiptrammell.com/static/discounting_for_patient_philanthropists.pdf
26. See, e.g., S. Scheffler, *The Rejection of Consequentialism*, Oxford, 1982.
27. That is to say, perhaps maximising consequentialism is false. See W. Sinnott-Armstrong, Consequentialism, *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*, https://plato.stanford.edu/archives/sum2019/entries/consequentialism/
28. Greaves and MacAskill (2019) discuss this claim in more detail.
29. This case was first presented by P. Singer, Famine, affluence, and morality, *Philosophy and Public Affairs* 1, no.3 (1972) 229-243.
30. That is unless the best short-term initiatives happen to have long-run benefits comparable to those of the longtermist interventions discussed above. This condition seems to us unlikely.

**3. Navigating the Next Century's Challenges**

1. The world's 2,095 billionaires are worth $8 trillion (Forbes Billionaires 2020. (n.d.). Accessed 25 September 2020. https://web.archive.org/save/https://www.forbes.com/billionaires/#b91420e251c7). In 2017, the most recent year we have data for, 689 million people lived

on less that $1.90 a day (PovcalNet: the on-line tool for poverty measurement developed by the Development Research Group of the World Bank. Accessed 25 September 2020. https://web.archive.org/save/http://iresearch.worldbank.org/PovcalNet/povDuplicateWB.aspx.)
2. M. Roser, H. Ritchie, and E. Ortiz-Ospina, World Population Growth, 2013, Accessed 12 October 2020.
3. M. Roser, Fertility Rate, 2019, Accessed 10 February 2020.
4. World Population Prospects 2019: Highlights, *Statistical Papers - United Nations (Ser. A), Population and Vital Statistics Report*.
5. The United Nations on World Population in 2300, *Population and Development Review* 30, no.1 (2004) 181-187.
6. J. Kasper and F. Bajunirwe, Brain Drain in Sub-Saharan Africa: Contributing Factors, Potential Remedies and the Role of Academic Medical Centres, *Archives of Disease in Childhood* 97, no. 11 (2012) 973-979.
7. Y. M. Bar-On, R. Phillips and R. Milo, The biomass distribution on Earth, 2018.
8. See IPCC. (2013). Climate Change 2013: The Physical Science Basis; J. Rogelj et al, Paris Agreement climate proposals need a boost to keep warming well below 2 °C, *Nature* 534 no. 7609 (2016) 631-639.
9. IPCC. (2018). Global warming of 1.5°C. SR15 summary for policymakers.
10. Alan Jacobs's argument that interest groups generally prefer distributive gains to intertemporal bargains. *Governing for the long term: Democracy and the politics of investment*, 2011.
11. S. Richard, Cuba embarks on a 100-year plan to protect itself from climate change, 2018, Web Archive. Accessed 25 September 2020.
12. I. Hadjipaschalis, A. Poullikkas and V. Efthimiou, Overview of current and future energy storage technologies for electric power applications, *Renewable and Sustainable Energy Reviews* 13, no. 6-7, 1513-1522.
13. For more on this, see R. Yampolskiy, *Unexplainability and Incomprehensibility of AI*, 2020.
14. Jacobs and Matthews, *Why Do Citizens Discount the Future?*
15. G. Brundtland, Report of the World Commission on environment and development: Our common future, *United Nations General Assembly* document A/42/427, 1987.

**4. Longtermist Institutional Reform**

1. We are grateful to Adam Gibbons, Alexander Guerrero, and Toby Ord for comments on previous drafts.
2. This argument for longtermism is made in much greater detail in H. Greaves and W. MacAskill, The case for strong longtermism, *Global Priorities Institute Working Paper Series*, *GPI Working Paper* 7 (2019) and in H. Greaves, W. MacAskill, E. Thornley, The case for strong longtermism, this volume.
3. S. A. Binder, Can congress legislate for the future? *John Brademas Center for the Study of Congress, New York University, Research Brief* 3 (2006); I. González-Ricoy and A. Gosseries, Designing institutions for future generations, *Institutions for Future Generations* (2016) 3-23.
4. *Central Intelligence Agency,* Field listing: budget, *The World Fact Book 2020,* Washington DC. https://www.cia.gov/library/publications/the-world-factbook/fields/224.html; International Monetary Fund, General government total expenditure, 2015-2022. *World Economic Outlook Database*, April 2017.
5. L. Berkowitz and N. Walker, Laws and moral judgments, *Sociometry* (1967) 410-422; K. Bilz and J. Nadler, Law, Psychology, and Morality, *Psychology of Learning and Motivation* 50 (2009) 101-131; A. R. Flores and S. Barclay, Backlash, consensus, legitimacy, or polarization: The effect of same-sex marriage policy on mass attitudes, *Political Research Quarterly* 69, no. 1 (2016) 43-56; M. E. Tankard and E. L. Paluck, Norm perception as a vehicle for social change, *Social Issues and Policy Review* 10, no. 1 (2016) 181-211; T. R. Tyler, *Why People Obey the Law*, 2006; N. Walker and M. Argyle, Does the law affect moral judgments? *The British Journal of Criminology* 4, no. 6 (1964) 570-581.
6. This typology follows S. Caney, Political institutions for the future: a five-fold package, *Institutions for Future Generations*, UK (2016) 135-155 and I. González-Ricoy and A. Gosseries, Design-

ing institutions for future generations, *Institutions for Future Generations*, UK (2016) 3-23.

7. S. Frederick, G. Loewenstein and T. O'Donoghue, Time discounting and time preference: a critical review, *Journal of Economic Literature* 40, no. 2 (2002) 351-401; González-Ricoy and Gosseries 2016; Y. Halevy, Strotz meets allais: diminishing impatience and the certainty effect, *American Economic Review* 98, no. 3 (2008) 1145-62; K.Irving, Overcoming short- termism: Mental time travel, delayed gratification and how not to discount the future, *Australian Accounting Review* 19, no. 4 (2009) 278-294; A. M. Jacobs and J. S. Matthews, why do citizens discount the future? Public opinion and the timing of policy consequences, *British Journal of Political Science* (2012) 903-935.
8. A. M. Jacobs, Policymaking for the long term in advanced democracies, *Annual Review of Political Science* 19 (2016) 433-454.
9. M. K. MacKenzie, Institutional design and sources of short-termism, *Institutions for Future Generations*, UK (2016) 24-48.
10. Whether this is the best model of rational longtermist decision-making is not a closed question. For some discussion, see: A. Askell, Evidence neutrality and the moral value of information, *Effective Altruism: Philosophical Issues*, UK (2019) 37-52; C. Tarsney, The epistemic challenge to longtermism, *Global Priorities Institute Working Paper Series, GPI Working Paper* 10 (2019); D. Thorstad and A. Mogensen, Heuristics for clueless agents: how to get away with ignoring what matters most in ordinary decision-making, *Global Priorities Institute Working Paper Series, GPI Working Paper* 2 (2020).
11. Caney 2016; E. U. Weber, Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet), *Climatic Change* 77, no. 1-2 (2006) 103-120.
12. Caney 2016; D. Johnson and S. Levin, The tragedy of cognition: psychological biases and environmental inaction, *Current Science* (2009) 1597.
13. A. M. Jacobs, Policymaking for the long term in advanced democracies, *Annual Review of Political Science* 19 (2016) 433-454.
14. J. Bidadanure, Youth quotas, diversity, and long-termism: can young people act as proxies for future generations? *Institutions for Future Generations,* UK (2016) 266-281; Frederick, Loewenstein and O'Donoghue, (2002) 351-401; M. K. MacKenzie (2016) 24-48.
15. A. Chrisoula, Environmental preservation and second-order procrastination, *Philosophy and Public Affairs* 35, no. 3 (2007) 233-248; A. Chrisoula and M. D. White, eds, *The Thief of Time: Philosophical Essays on Procrastination*, 2010; N. J. Stroud, Polarization and partisan selective exposure, *Journal of Communication* 60, no. 3 (2010) 556-576, 51-67.
16. Caney, Political institutions for the future, 135-155.
17. A. R. Douglas, *The Logic of Congressional Action*, 1990; Binder 2006; Caney 2016; D. R. Mayhew, *Congress: The Electoral Connection*, 1974; E. R. Tufte, *Political Control of the Economy*, 1978; For a contrary view, see N. Beck, Does there exist a political business cycle: a Box-Tiao analysis, *Public Choice* 38, no. 2 (1982) 205-209.
18. Caney 2016.
19. Binder 2006.
20. A. Alberto and G. Tabellini, Credibility and politics, *European Economic Review* 32, no. 2-3 (1988) 542-550; A. M. Jacobs, Policymaking for the long term in advanced democracies, *Annual Review of Political Science* 19 (2016) 433-454; T. Persson and G. E. Tabellini, *Monetary and Fiscal Policy. Vol. 1, Credibility*, 1994.
21. V. A. Chanley, T. J. Rudolph, and W. M. Rahn, The origins and consequences of public trust in government: a time series analysis, *Public Opinion Quarterly* 64, no. 3 (2000) 239-256; J. P. Clinch and L. Dunne, Environmental tax reform: an assessment of social responses in Ireland, *Energy Policy* 34, no. 8 (2006) 950-959; M. J. Hetherington, *Why Trust Matters: Declining Political Trust and the Demise of American Liberalism*, 2005; B. Simonsen and M. D. Robbins, Reasonableness, satisfaction, and willingness to pay property taxes, *Urban Affairs Review* 38, no. 6 (2003) 831-854.
22. This subsection owes a considerable debt to Caney 2016.
23. N. Jones, M. O'Brien and T. Ryan, Representation of future generations in United Kingdom policymaking, *Futures* 102 (2018) 153-163.
24. B. Bimber, Congressional support agency products and services for science and technology issues: A survey of congressional staff attitudes about the work of CBO, CRS, GAO, and OTA, Paper prepared for the *Carnegie Commission on Science, Technology, and Government,* 1990; J.

Warner and G. Tudor, The Congressional Futures Office, Paper prepared for the *Belfer Center for Science and International Affairs, Harvard Kennedy School*, May 2019.
25. Binder 2006.
26. Tudor and Warner 2019.
27. M. Coleman, L. Devaney, D. Torney, and P. Brereton, Ireland's world-leading citizens' climate assembly, what worked? What didn't? *Climate Home News*, 27 June, 2019, https://www.climatechangenews.com/2019/06/27/irelands-world-leading-citizens-climate-assembly-worked-didnt/
28. At the time of writing, these deliberations are not yet complete.
29. There is some empirical evidence for this hypothesis in the literature on Demeny voting: A. Reiko and R. Vaithianathan, Intergenerational voter preference survey-preliminary results, *Hitotsubashi University Repository*, 2012; as well as in the literature on sociological institutionalism: R. E. Goodin, *Utilitarianism as a Public Philosophy*, 1995.
30. J. S. Fishkin and R. C. Luskin, Experimenting with a democratic ideal: Deliberative polling and public opinion, *Acta Politica* 40, no. 3 (2005) 284-298; J. S. Fishkin, R. W. Mayega, L. Atuyambe, N. Tumuhamye, J. Ssentongo, A. Siu, and W. Bazeyo, Applying deliberative democracy in Africa: Uganda's first deliberative polls, *Daedalus* 146, no. 3 (2017) 140-154; C. List, R. C. Luskin, J. S. Fishkin and I. McLean, Deliberation, single-peakedness, and the possibility of meaningful democracy: Evidence from deliberative polls. *The Journal of Politics* 75, no. 1 (2013) 80-95.
31. Code of Federal Regulations 32 U.S.C. § 651.42
32. Gov.UK, Guidance: Environmental Impact Assessments, 6 March 2014, https://www.gov.uk/guidance/environmental-impact-assessment
33. California Public Resources Code § 21000 et seq.
34. Friends of the Oldman River Society *v.* Canada (Minister of Transport) [1992] 1 S.C.R. 3 (1992).
35. *Well-being of Future Generations Bill*, HL Bill 15, 2019-2020.
36. Jones, O'Brien, and Ryan 2018.
37. E. Dal Bó and M. Rossi. *Term Length and Political Performance*, no. w14511, National Bureau of Economic Research, 2008; E. Dal Bó and M. Rossi, Term length and the effort of politicians, *The Review of Economic Studies* 78, no. 4 (2011) 1237-1263; R. Titiunik and A. Feher, Legislative behaviour absent re- election incentives: findings from a natural experiment in the Arkansas Senate, *Journal of the Royal Statistical Society: Series A*, 181(2) (2018): 351-378.
38. R. Aoki and R. Vaithianathan, Intergenerational voter preference survey-preliminary results, *Hitotsubashi University Repository,* 2012; Y. Kamijo, T. Tamura, and Y. Hizen, Effect of proxy voting for children under the voting age on parental altruism towards future generations, *Futures* 122, (2020); Vaithiamathan *et al*. 2013.

### 5. Lessons from the British Welfare State for Future Generations Legislation

1. For a detailed account of working class life in Britain, please see H. David, *Working Lives: the Forgotten Voices of Britain's Post-War Working Class*, 2014.
2. See S. Gorard, and N. Siddiqui, Grammar schools in England: a new analysis of social segregation and academic outcomes, *British Journal of Sociology of Education* 39, no. 7 (2018) 909-924. and S. Themelis, Meritocracy through education and social mobility in post- war Britain: a critical examination. *British Journal of Sociology of Education* 29, no. 5 (2008): 427-38.
3. See HM Government, An evidence review of the drivers of child poverty for families in poverty now and for poor children growing up to be poor adults (2014).
4. See S. Fitzpatrick, G. Bramley, and S. Johnsen, Pathways into multiple exclusion homelessness in seven UK cities, *Urban Studies* 50, no. 1 (2013).

### 6. The Challenge of Effective Long-term Thinking in the UK Government and the Critical Role of Philanthropy

1. Department for Business, Innovation and Skills, A Long Term Focus for Corporate Britain, Discussion Paper, URN 10/1225 (2010) http://www.bis.gov.uk/assets/biscore/business-law/docs/1/10-1225-long-term-focus-corporate-britain ; W. Lazonick, Profits without prosperity,

*Harvard Business Review* 92 (2014) 46-55; L. Martin, Takeover bids in the target's boardroom, *The Business Lawyer*, 1979, 101-134.
2. Institute for Government, 'Civil service staff numbers (FTE) over time', 2010, https://www.instituteforgovernment.org.uk/charts/civil-service-staff-numbers-fte-over-time
3. OECD, *Enhancing productivity in UK core cities: connecting local and regional growth*, OECD Urban Policy Reviews, OECD Publishing, Paris, 2020, https://doi.org/10.1787/9ef55ff7-en
4. UK2070 Commission, *Make No Little Plans: Acting at Scale for a Fairer and Stronger Future*, 2020, http://uk2070.org.uk/wp-content/uploads/2020/02/UK2070-FINAL-REPORT.pdf
5. Foresight Flood and Coastal Defence Project, *Foresight Future Flooding Executive Summary*, 2004, online https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/300332/04-947-flooding-summary.pdf
6. George Peabody quoted by the Peabody Historical Society.
7. The Letters of George Peabody.
8. From correspondence of March 1862 and January 1866.

### 7. A Little Bit of Funding Goes a Long Way: The APPG for Future Generations

1. More information on the APPG for Future Generations is available on their website at www.appgfuturegenerations.com.
2. N. Jones, M. O'Brien, and T. Ryan, Representation of future generations in United Kingdom policymaking, *Futures* 102 (2018) 153-163. This paper along with some of our other research is accessible at www.appgfuturegenerations.com/research.
3. See Bird, *Lessons from the British Welfare State for future generations legislation*, this volume.
4. Alpenglow is a non profit which aims to put long-term thinking at the heart of UK policymaking. Please see www.alpenglow.org.uk for further information.

### 8. Biosecurity, Longtermism, and Global Catastrophic Biological Risks

1. M. Schoch-Spana *et al.* Global Catastrophic Biological Risks: toward a working definition, *Health Security* 15. no. 4, https://www.liebertpub.com/doi/full/10.1089/hs.2017.0038
2. *J. Yassif*, *Reducing Global Catastrophic Biological Risks*, *Health Security 15. no. 4, https://www.liebertpub.com/doi/full/10.1089/hs.2017.0038*
3. 80,000 Hours Podcast: Dr. Cassidy Nelson on the twelve best ways to stop the next pandemic (and limit COVID-19), 2020, https://80000hours.org/podcast/episodes/cassidy-nelson-12-ways-to-stop-pandemics/
4. D. Carrington, Coronavirus: 'Nature is sending us a message', says UN environment chief, The Guardian, March 25, 2020.
5. *N. Bostrom*, *The Vulnerable World Hypothesis*, *Global Policy 10, issue 4 (2019) 466.*
6. The one potential downside is that this is a double-edged sword. The COVID-19 pandemic might also provide valuable insights to malicious actors about the potential impact of a deliberately caused biological event. The biosecurity community is still thinking through the potential impacts of this event on the evolving risk landscape.
7. The Overton Window is a model for understanding how ideas in society change over time and influence politics. It was developed in the 1990s by Joseph P. Overton. See: https://www.mackinac.org/OvertonWindow
8. The US Centers for Disease Control define the case fatality rate as follows: "…the proportion of persons with a particular condition (cases) who die from that condition. It is a measure of the severity of the condition. The formula is: Number of cause-specific deaths among the incident cases divided by the total number of incident cases." See: https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section3.html
9. E. Cameron, R. Katz, J. Konyndyk and M. Nalabandian, A spreading plague: Lessons and recommendations for responding to a deliberate biological event, *NTI Paper*, June 2019. https://media.nti.org/documents/NTI_Paper_A_Spreading_Plague_FINAL_061119.pdf
10. *Jeremy Konyndyk at the Center for Global Development is leading a research project in partnership*

*with NTI | bio on Operational Response Preparedness for High-Consequence Biological Events. The goal of this work is to develop actionable recommendations regarding response strategies for large-scale biological events, to enable national leaders and international organisations to marshal an effective, anticipatory response to pandemics like COVID-19, and other potentially more severe biological events, that may emerge in the future. This work will be published in early 2021.*

11. While there is some uncertainty among longtermism-focused communities regarding the effectiveness of response preparedness for GCBRs, we at NTI, along with many biosecurity and pandemic preparedness colleagues, are of the view that it is in fact a critically important and effective tool for meaningfully reducing these risks.
12. *A. Sandberg and C. Nelson, Who should we fear more: biohackers, disgruntled postdocs, or bad governments? A simple risk chain model of biorisk, Health Security 18,* no. 3 (2020).
13. *R. Danzig et al. Aum Shinrikyo: insights into how terrorists develop chemical and biological weapons, 2012, https://www.cnas.org/publications/reports/aum-shinrikyo-insights-into-how-terrorists-develop-biological-and-chemical-weapons*
14. S. R. Carter, S. S. Morse, J. M. Yassif, Proposed global norms for microbiology, synthetic biology, and emerging biotechnologies, *NTI Biosecurity Innovation and Risk Reduction Initiative*, 2019, https://media.nti.org/documents/Framing_Paper_-_Norms_for_Microbiology_Synthetic_Biology_and_Emerging_Technolo_BzcMHsF.pdf )
15. *I. Nath, J. M. Yassif, Establishing a seal of approval to incentivize adherence to biosecurity norms, NTI Biosecurity Innovation and Risk Reduction Initiative, 2018, https://media.nti.org/documents/NTI_Paper_5_Establishing_a_Seal_of_Approval_to_Incentivize_Adherence_to_Biosec_a5iL1rO.pdf)*
16. *M. J. Palmer, S. M. Hurtley, S. W. Evans, Visibility initiative for responsible science, paper presented at the NTI Biosecurity Innovation and Risk Reduction Initiative meeting, Geneva, Switzerland, 15 October 2019.*
17. *NTI and World Economic Forum, Biosecurity innovation and risk reduction: a global framework for accessible, safe and secure DNA synthesis, 2020, Available at https://wwwweforum.org/reports/biosecurity-innovation-and-risk-reduction-a-global-framework-for-accessible-safe-and-secure-dna-synthesis-582d582cd4*
18. *D. Friedman, One percent: instituting biosecurity investment, NTI Biosecurity Innovation and Risk Reduction Initiative, 2018, https://media.nti.org/documents/NTI_Initiative_- _Paper_2_-_Instituting_Biosecurity_Investment.pdf*
19. *N. Bostrom, Information hazards: a typology of potential harms from knowledge, Review of Contemporary Philosophy 10* (2011) *44-79.*
20. *B. Cameron, J. M. Yassif, J. Jordan and J. Eckles, preventing global catastrophic biological risks: lessons and recommendations from a tabletop exercise held at the 2020 Munich Security Conference, NTI | bio, September 2020. Available at: https://media.nti.org/documents/NTI_BIO_TTX_RPT_FINAL.pdf*
21. *Toby Ord's book, The Precipice: Existential Risk and the Future of Humanity*, 2020, *is a great example of work to explain longtermist views and concepts in a way that is accessible to the other communities.*
22. Johns Hopkins University Center for Health Security, Clade X Exercise, 2018. This exercise presented a fictional GCBR scenario involving a terrorist attack with an engineered pathogen. https://www.centerforhealthsecurity.org/our-work/events/2018_clade_x_exercise/
23. Emerging Leaders in Biosecurity Fellowship. https://www.centerforhealthsecurity.org/our-work/emergingbioleaders/
24. 80,000 Hours. https://80000hours.org/

**9. Utilising Insurance for Climate Risk Reduction in the UK**

1. M. Carney, *Breaking the tragedy of the horizon: climate change and financial stability*, 2015, https://www.bankofengland.co.uk/speech/2015/breaking-the-tragedy-of-the-horizon-climate-change-and-financial-stability.
2. UNISDR, *Economic Losses, Poverty and Disasters: 1998-2017,* 2018 https://www.unisdr.org/2016/iddr/IDDR2018_Economic%20Losses.pdf.
3. M. Carney, *Remarks given during the UN Secretary General's Climate Action Summit 2019. https://www.bankofengland.co.uk/speech/2019/mark-carney-speech.*
4. Cabinet Office, *National Risk Register,* 2020, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/61934/national_risk_register.pdf
5. Home Office, Department of Health and Social Care and Department for Environment, Food

and Rural Affairs, *UK Biological Security Strategy,* 2018. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/730213/2018_UK_Biological_Security_Strategy.pdf

6. HM Treasury, *Government as insurer of last resort: managing contingent liabilities in the public sector,* 2020, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/871660/06022020_Government_as_Insurer_of_Last_Resort_report__Final_clean_.pdf
7. Home Office, *CONTEST: The United Kingdom's strategy for countering terrorism,* 2011, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/716907/140618_CCS207_CCS0218929798-1_CONTEST_3.0_WEB.pdf
8. Committee on Climate Change, *Reducing UK emissions: 2020 Progress Report to Parliament,* https://www.theccc.org.uk/publication/reducing-uk-emissions-2020-progress-report-to-parliament/
9. T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, London, 2020.
10. See also John and MacAskill, *Longtermist Institutional Reform*, this volume.
11. P. Jarzabkowski, K. Chalkias, D. Clarke, E. Iyahen, D. Stadtmueller and A. Zwick, *Insurance for climate adaptation: opportunities and limitations,* 2019.
12. K. Schanz, *Underinsurance in mature economies: reasons and remedies,* 2019, https://www.genevaassociation.org/research-topics/socio-economic-resilience/underinsurance-mature-economies-reasons-and-remedies
13. https://www.jbs.cam.ac.uk/wp-content/uploads/2020/08/crs-lloydscityriskindex-execsummary.pdf
14. A. Levite, S. Kannry, W. Hoffman, *Addressing the private sector cybersecurity predicament: the indispensable role of insurance,* 2018, https://www.jstor.org/stable/resrep20984.
15. P. Jarzabkowski, K. Chalkias, D. Clarke, E. Iyahen, D. Stadtmueller and A. Zwick, *Insurance for climate adaptation: Opportunities and limitations,* 2019.
16. R. Smith-Bingham, A. Wittenburg, *Building national resilience: Aligning mindsets, capabilities, and investment*, 2020, https://www.jstor.org/stable/resrep20984.
17. HM Treasury. (2020) *Government as insurer of last resort: managing contingent liabilities in the public sector.*
18. T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, London, 2020.
19. OASIS, *Ground-breaking open source platform to be used for catastrophe modelling in the Philippines and Bangladesh,* 2018, https://www.insuresilience.org/ground-breaking-open-source- platform-to-be-used-for-catastrophe-modelling-in-the-philippines-and-bangladesh/.
20. P. Jarzabkowski, K. Chalkias, D. Clarke, E. Iyahen, D. Stadtmueller and A. Zwick. *Insurance for climate adaptation: Opportunities and limitations,* Washington DC, 2019, www.gca.org.
21. Swiss Re, *Successful Kenya livestock insurance program scheme scales up,* 2018, https://www.swissre.com/our-business/public- sector-solutions/thought-leadership/successful-kenya-livestock- insurance-program-scheme.html.
22. Dalberg, Impact evaluation of the R4 rural resilience initiative in Senegal, final report, *Oxfam and World Food Programme (WFP),* 2016, https://www.oxfamamerica.org/static/media/ files/WFP_Oxfam_R4_Final_Report_English_FINAL.pdf.
23. Oxford Policy Management. *Independent evaluation of the African Risk Capacity,* 2017, https://www. opml.co.uk/projects/independent-evaluation-african-risk-capacity
24. P. Jarzabkowski, K. Chalkias, D. Clarke, E. Iyahen, D. Stadtmueller & A. Zwick, *Insurance for climate adaptation: opportunities and limitations,* 2019.
25. P. Jarzabkowski, K. Chalkias, D. Clarke, E. Iyahen, D. Stadtmueller & A. Zwick, *Insurance for climate adaptation: opportunities and limitations,* 2019.
26. United Nations Office for Disaster Risk Reduction (UNDRR), Words into Action guidelines: build back better in recovery, rehabilitation and reconstruction, *United Nations Office for Disaster Risk Reduction (UNDRR)*, 2017, https://www.unisdr.org/we/inform/ publications/53213.
27. All Party Parliamentary Group for Future Generations, https://www.appgfuturegenerations.com/
28. https://publications.parliament.uk/pa/ld5801/ldselect/ldliaison/102/10205.htm

**10. Ensuring the Safety of Artificial Intelligence**

1. Section 5 of this chapter is based on the work of A. Askell, M. Brundage and G. Hadfield, The role of cooperation in responsible AI development, 2019. Thanks to Jack Clark for helpful comments.

2. This work was primarily completed while Dr. Askell was a Research Scientist in Policy at OpenAI.
3. G. Easterbrook, Forgotten benefactor of humanity, *The Atlantic*, 1997. https://www.theatlantic.com/magazine/archive/1997/01/forgotten-benefactor-of-humanity/306101/
4. D. Holloway, Nuclear weapons and the escalation of the Cold War, 1945-1962, *The Cambridge History of the Cold War*, 376-397, Cambridge, 2010; T. Ord, *The Precipice: Existential Risks and the Future of Humanity*, 2020, Chapter 4.
5. Some have even argued that AI has the characteristics of a general purpose technology akin to electricity and computers. See: E. Brynjolfsson, D. Rock and C. Syverson, Artificial intelligence and the modern productivity paradox, *The Economics of Artificial Intelligence: An Agenda*, 23 (2019).
6. S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th Edition.
7. The paradox is expressed in H. Moravec, Mind Children, 1988.
8. J. Lighthill, Artificial intelligence: A general survey, *Artificial Intelligence: a paper symposium* (1973) 1-21, London.
9. N. J. Nilsson, *The Quest for Artificial Intelligence*, 2009; Human-Level Artificial Intelligence? Be Serious! *AI Magazine* 26, no.4 (2005) 68, https://doi.org/10.1609/aimag.v26i4.1850
10. Artificial neurons are often not closely modeled on biological neurons. However some artificial neurons, such as those in spiking neural networks, do attempt to more closely model biological neurons: A.Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier and A. Maida, Deep learning in spiking neural networks, *Neural Networks*, *111* (2019) 47-63.
11. Y. LeCun, Y. Bengio and G. Hinton, Deep learning, *Nature* 521 (2015) 436-444, https://doi.org/10.1038/nature14539 provides an overview of deep learning and its innovations in areas like images, video, and text. As Y. Bengio in System 1 deep learning to system 2 deep learning, *Posner lecture at NeurIPS 2019* highlights, neural networks are also often worse at the kind of system 2 thinking at which symbolic systems excel. For example, contemporary language models that can produce coherent, novel text can still struggle with abstract tasks like identifying entailments or performing three or four digit addition, as shown in T. B. Brown, B. Mann, N. Ryder, M. Subbiah, Language models are few-shot learners, 2020.
12. D. Silver, J. Schrittwieser, K. Simonyan, Mastering the game of Go without human knowledge, *Nature* 550 (2017) 354-359, https://doi.org/10.1038/nature24270
13. The GPT-3 model achieved near state of the art in CoQA (S. Reddy, D. Chen and C. D. Manning, Coqa: A conversational question answering challenge, *Transactions of the Association for Computational Linguistics*, 7 (2019) 249-266) and above the fine-tuned state of the art on TriviaQA (M. Joshi, E. Choi, D. S. Weld and L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017) for example, without being fine-tuned for these tasks (Mann, Ryder & Subbiah et al., 2020).
14. D. Amodei and D. Herndandez, AI and Compute, OpenAI, 2018, https://openai.com/blog/ai-and-compute/
15. J. Shalf, The future of computing beyond Moore's law, *Philosophical Transactions of the Royal Society A*, 2020. For an overview of the role of computer chips in AI see: S. M. Khan and A. Mann, AI Chips: What They Are and Why They Matter, *Center for Security and Emerging Technology*, April 2020, cset.georgetown.edu/research/ai-chips-what-they-are-and-why-they-matter/
16. See: Silver, Schrittwieser, Simonyan (2017). A non-robotic AI system may not be capable of doing all the non-intellectual work that humans currently do, such as physical tasks. The degree to which embodiment may be an essential component in the training of AI is an open question, however. See: M. L. Anderson, Embodied cognition: A field guide, *Artificial Intelligence* 149, no.1 (2003), 91-130; T. Ziemke, What's that thing called embodiment? *Proceedings of the annual meeting of the cognitive science society* 25 no. 25 (2003).
17. R. Gruetzemacher, D. Paradice and K. B. Lee, Forecasting extreme labor displacement: A survey of AI practitioners, *Technological Forecasting and Social Change* 161 (2020); K. Grace, J. Salvatier, A. Dafoe, B. Zhang and O. Evans, When will AI exceed human performance? Evidence from AI experts, *Journal of Artificial Intelligence Research* 62 (2018) 729-754 offer recent surveys of AI practitioners on the potential for AI labor displacement.
18. *17* M. Webb, The impact of artificial intelligence on the labor market, 2019, Available at SSRN: https://ssrn.com/abstract=3482150 or http://dx.doi.org/10.2139/ssrn.3482150 and M. Muro, J. Whiton and R. Maxim, What jobs are affected by AI? *Brookings Institute*, 2019, https://www.brookings.edu/wp-content/uploads/2019/11/2019.11.20_BrookingsMetro_What-jobs-are-affected-by-

AI_Report_Muro-Whiton-Maxim.pdf both note that high-skilled jobs may be particularly exposed to AI. This is in contrast with other forms of automation that tend to affect low-skilled work.

19. There is sometimes thought to be a conflict between those focused on 'near term' and 'long term' impact of AI. Although this chapter is explicitly focused on long-term consequences of AI, this isn't intended to be in conflict with work on immediate impacts from AI. For a critical review of the division of work into 'near term' and 'long term' see: C. Prunkl, and J. Whittlestone, Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, 138-14.
20. C. O'Keefe, P. Cihon, B. Garfinkel, C. Flynn, J. Leung and A. Dafoe, The Windfall Clause: *Distributing the Benefits of AI for the Common Good* (2020) 327-331 and M. Brundage and S. Avin, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation, 2018.
21. Brundage & Avin et al., 2018.
22. R. Zwetsloot and A. Dafoe, Thinking about risks from AI: accidents, misuse and structure, *Lawfare*, 2019 https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure
23. Another reason is that we often discount the welfare of individuals that exist in the future. The wisdom of pure temporal discounting has been questioned by T. Cowen and D. Parfit, Against the social discount rate. Justice between age groups and generations, 1992, 144-145. I will not focus on it here.
24. The value of information is higher in high-uncertainty domains. In Askell, 2019 I argue that it can even be worth investing in more speculative interventions, all else being equal, because the value of the information we get from doing so is higher.
25. R. J. Lempert and M. T. Collins, Managing the risk of uncertain threshold responses: comparison of robust, optimum, and precautionary approaches, *Risk Analysis: An International Journal* 27, no.4 (2007) 1009-1026; R. J. Lempert, D. G. Groves, S. W. Popper and S. C. Bankes, A general, analytic method for generating robust strategies and narrative scenarios, *Management Science* 52, no.4 (2006) 514-528.
26. It is not possible to give an overview of these various benchmarks here. They include, for example: Y. LeCun, C. Cortes and C. Burges, MNIST handwritten digit database, 2010, http://yann.lecun.com/exdb/mnist/ ; K. Papineni, S. Roukos, T. Ward and W. J. Zhu, BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, 311-318; P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, Squad: 100,000+ questions for machine comprehension of text, 2016; A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S. R. Bowman, Glue: A multi-task benchmark and analysis platform for natural language understanding, 2018; SuperGLUE, A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill and S. Bowman, Superglue: A stickier benchmark for general-purpose language understanding systems, *Advances in Neural Information Processing Systems*, 2019, 3266-3280; Facebook AI, Introducing Dynabench: Rethinking the way we benchmark AI, 2020, https://ai.facebook.com/blog/dynabench-rethinking-ai-benchmarking/
27. M. R. Frank, D. Autor, J. E. Bessen, E. Brynjolfsson, M. Cebrian, D. J. Deming and D. Wang, Toward understanding the impact of artificial intelligence on labor, *Proceedings of the National Academy of Sciences* 116, no. 14 (2019) 6531-6539.
28. Section 5 of M. Muro, R. Maxim, and J. Whiton, How machines are affecting people and places, *Brookings Institute*, 2019 https://www.brookings.edu/wp-content/uploads/2019/01/2019.01_BrookingsMetro_Automation-AI_Report_Muro-Maxim-Whiton-FINAL-version.pdf discusses some of the difficulties involved in adjusting to automation and steps that can be taken to mitigate them.
29. As D. Acemoglu and P. Restrepo, Artificial intelligence, automation and work, *National Bureau of Economic Research*, no. w24196, 2018, point out, automation occurs at the level of tasks rather than at the level of occupation. AI could partially or fully automate tasks that are common to different occupations.
30. S. Mishra, J. Clark, and C. R. Perrault, *Measurement in AI Policy: Opportunities and Challenges*, 2020.
31. NIST — The National Institute of Standards and Technology, AI Research — Foundational, 2020, https://www.nist.gov/topics/artificial-intelligence/ai-research-foundational
32. S. Russell, *Human compatible: Artificial intelligence and the problem of control*, 2019.
33. W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, Definitions, methods, and applications in interpretable machine learning, *Proceedings of the National Academy of Sciences* 2019, no. 44 (2019) 22071-22080.

34. This is not to suggest that there is nothing we can do to reduce bias in language models. Progress may be made by filtering or tagging training data, adding de-biasing text to the context, building bias metrics for language models, and various other methods. For example, S. Bordia and S. R. Bowman, Identifying and reducing gender bias in word-level language models, 2019 recently developed a method for reducing gender bias in language models. The point is that the task of reducing unwanted outputs in generative language models is not a trivial one.
35. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman and D. Mané, Concrete problems in AI safety, 2016.
36. Not all work related to AI safety and alignment, as they are broadly defined above, is carried out by AI safety researchers, however.
37. G. Irving, P. Christiano and D. Amodei, AI safety via debate, 2018; G. Irving and A. Askell, AI safety needs social scientists, *Distill* 4, no. 2 (2019) e14, https://distill.pub/2019/safety-needs-social-scientists/
38. Other research programs in AI safety include cooperative inverse reinforcement learning (See: D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, Cooperative inverse reinforcement learning, *Advances in neural information processing systems*, 2016, 3909-3917.) and recursive reward modeling (See: J. Leike, D. Krueger, T. Everitt, M. Martic, V. Maini and S. Legg, Scalable agent alignment via reward modeling: a research direction, 2018.), among others.
39. This includes theoretical safety work, practical safety work, and work to clarify and expand on the problems and solutions already identified. Ought (ought.org) is an example of a non-profit organization that's doing AI safety-related research, while the Stanford Center for AI Safety (aisafety.stanford.edu) and the UC Berkeley Center for Human-Compatible Artificial intelligence (humancompatible.ai) are examples of an academic institutions doing safety-related research.
40. I. Solaiman, M. Brundage, J. Clark, A. Askell, A. Herbert-Voss, J. Wu and M. McCain, *Release strategies and the social impacts of language models*, 2019.
41. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson and T. Gebru, Model cards for model reporting, *In Proceedings of the conference on fairness, accountability, and transparency* (2019) 220-229.
42. C. O'Keefe, P. Cihon, B. Garfinkel, C. Flynn, J. Leung and A. Dafoe, The Windfall Clause: Distributing the Benefits of AI for the Common Good, 2020, 327-331.
43. J. Clark, G. K. Hadfield, Regulatory Markets for AI Safety, 2019.
44. As we will see, they also face collective action problems involving other regulators.
45. See F. Suarez and G. Lanzolla, The half-truth of first-mover advantage, *Harvard Business Review*, 2005 on how first-mover advantages can be overstated and W. Hunt, The Flight to Safety-Critical AI: Lessons in AI Safety from the Aviation Industry, *CLTC White Paper Series, UC Berkeley Center for Long-Term Cybersecurity*, 2020 https://cltc.berkeley.edu/wp-content/uploads/2020/08/Flight-to-Safety-Critical-AI.pdf who argues that the empirical evidence of a race to the bottom is limited. However, Hunt (2020) notes that regulation has played a role in making it hard for those in the aviation industry to cut corners on AI safety.
46. Askell, Brundage and Hadfield, 2019; S. Armstrong, N. Bostrom and C. Shulman, Racing to the precipice: a model of artificial intelligence development, AI & society 31, no.2 (2019) 201-206.
47. Askell, Brundage and Hadfield, 2019.
48. M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, and T. Maharaj, Toward trustworthy AI development: mechanisms for supporting verifiable claims, 2020.
49. For example, see: H. Karnofsky, H, Potential Risks from Advanced Artificial Intelligence: The Philanthropic Opportunity, *Open Philanthropy*, 2016, https://www.openphilanthropy.org/blog/potential-risks-advanced-artificial-intelligence-philanthropic-opportunity ; A. Dafoe, A, AI governance: a research agenda, *Governance of AI Program, Future of Humanity Institute*, Oxford, UK, 2018; M. Brundage, Guide to working in artificial intelligence policy and strategy, *80,000 Hours*, 2019, https://80000hours.org/articles/ai-policy-guide/

### 11. Traversing the Garden of Forking Paths More Wisely

1. I want to thank Tyler John for research assistance in making this chapter as well as he and Will Fenning for useful comments to an earlier version of this piece. I gratefully acknowledge the

support by Open Philanthropy in the making of this article.

2. T. Edward, *Why Things Bite Back: Technology and the Revenge of Unintended Consequences,* New York, 1996.
3. I. Goldin and M. Mariathasan, *The Butterfly Defect: How Globalization Creates Systemic Risks, and What to Do about It*, Princeton, 2016.
4. T. Ord, *The Precipice. Existential Risk and the Future of Humanity*, New York, 2020.
5. S. Caney, Political institutions for the future: A fivefold package, *Institutions for Future Generations*, Oxford, 2016, 135–155.
6. I. González Ricoy and A. Gosseries, *Institutions for Future Generations*, Oxford, 2016.
7. *I speak on this occasion mainly about political decision-making because that is the realm I know best. I surmise that most of the issues I touch upon in this* chapter *are of relevance in other walks of life as well.*
8. M. F. Oppenheimer, *Pivotal Countries, Alternate Futures: Using Scenarios to Manage American Strategy*, Oxford and New York, 2016.
9. D. Kahneman, *Thinking, Fast and Slow*, London, 2011.
10. I. Hansson and C. Stuart, Malthusian Selection of Preferences, *The American Economic Review* 80, no. 3 (1990) 529–544.
11. T. Ord, *The Precipice. Existential Risk and the Future of Humanity*, New York, 2020.
12. H. A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization*. New York, 1959.
13. A. S. Whiting, The scholar and the policymaker. *World Politics*, 24, Supplement *Theory and Policy in International Relations*, 1972, 229–24.
14. A. Mintz and K. Jr. DeRouen, *Understanding Foreign Policy Decision Making*, New York, 2010.
15. G. Allison and P. Zelikow, *The Essence of Decision: Explaining the Cuban Missile Crisis*, Second Edition, New York, 1999.
16. D. J. Blake, Thinking ahead: government time horizons and the legalization of international investment agreements, *International Organization* 67, no. 4 (2013) 797–827.
17. See Caney (2016) who makes a forceful argument to expand the Finnish model to a much more ambitious, binding and widely adopted political practice.
18. V. Koskimaa and T. Raunio, Encouraging a longer time horizon: the Committee for the Future in the Finnish Eduskunta, *The Journal of Legislative Studies* 26, no.2, (2020) 159–179.
19. T. Carothers, Is democracy the problem? *The American Interest*, January 16, 2019, available at https://www.the-american-interest.com/2019/01/16/is-democracy-the-problem/
20. N. Gowing and C. Langdon, *Thinking the Unthinkable: A New Imperative for Leadership in a Disruptive Age*, 2018
21. *D*. Karpf, The 10,000-Year Clock Is a Waste of Time, *Wired*, January 29 2020, available at https://www.wired.com/story/the-10000-year-clock-is-a-waste-of-time/.
22. This idea has been adapted and expanded from N. Thompson who proposes that technologists should learn responsibility from the original iron ring; N. Thompson, It's time to push tech forward, and rebuild what it broke, *Wired*, October 15 2019, available at https://www.wired.com/story/wired25-work-together-fix-mess-we-made/.
23. For discussion of avoiding 'lock-in', see Chapter 5 of Ord (2020).