What We Owe The Future

William MacAskill

Contents

For my parents, Mair and Robin, and their parents, Ena and Tom and Daphne and Frank, and...

BASIC
BOOKS

PART I

THE LONG VIEW

Introduction

Imagine living, in order of birth, through the life of every human being who has ever lived.[1] Your first life begins about three hundred thousand years ago in Africa.[2] After living that life and dying, you travel back in time and are reincarnated as the second-ever person, born slightly later than the first. Once that second person dies, you are reincarnated as the third person, then the fourth, and so on. One hundred billion lives later,[3] you become the youngest person alive today. Your "life" consists of all of these lifetimes, lived consecutively.

Your experience of history is very different from what is depicted in most textbooks. Famous figures like Cleopatra or Napoleon account for a tiny fraction of your experience. The substance of your life is instead composed of ordinary lives, filled with everyday realities—eating, working, and socialising; laughing, worrying, and praying.

Your life lasts for almost four trillion years in total. For a tenth of that time, you're a hunter-gatherer, and for 60 percent you're an agriculturalist.[4] You spend a full 20 percent of your life raising children, a further 20 percent farming, and almost 2 percent taking part in religious rituals. For over 1 percent of your life you are afflicted with malaria or smallpox. You spend 1.5 billion years having sex and 250 million giving birth. You drink forty-four trillion cups of coffee.[5]

You experience cruelty and kindness from both sides. As a colonizer, you invade new lands; as the colonized, you suffer your lands taken from you. You feel the rage of the abuser and the pain of the abused. For about 10 percent of your life you are a slaveholder; for about the same length of time, you are enslaved.[6]

You experience, firsthand, just how unusual the modern era is. Because of dramatic population growth, a full third of your

life comes after AD 1200 and a quarter after 1750. At that point, technology and society begin to change far faster than ever before. You invent steam engines, factories, and electricity. You live through revolutions in science, the most deadly wars in history,[7] and dramatic environmental destruction. Each life lasts longer, and you enjoy luxuries that you could not sample even in your past lives as kings and queens. You spend 150 years in space and one week walking on the moon. Fifteen percent of your experience is of people alive today.[8]

That's your life so far—from the birth of *Homo sapiens* until the present. But now imagine that you live all future lives, too. Your life, we hope, would be just beginning. Even if humanity lasts only as long as the typical mammalian species (one million years), and even if the world population falls to a tenth of its current size, 99.5 percent of your life would still be ahead of you.[9] On the scale of a typical human life, you in the present would be just five months old. And if humanity survived longer than a typical mammalian species—for the hundreds of millions of years remaining until the earth is no longer habitable, or the tens of trillions remaining until the last stars burn out—your four trillion years of life would be like the first blinking seconds out of the womb.[10] The future is big.

If you knew you were going to live all these future lives, what would you hope we do in the present? How much carbon dioxide would you want us to emit into the atmosphere? How much would you want us to invest in research and education? How careful would you want us to be with new technologies that could destroy or permanently derail your future? How much attention would you want us to give to the impact of today's actions on the long term?

I present this thought experiment because morality, in central part, is about putting ourselves in others' shoes and treating their interests as we do our own. When we do this at the full scale of human history, the future—where almost everyone lives and where almost all potential for joy and misery lies—comes to the fore.

This book is about *longtermism*: the idea that positively influencing the longterm future is a key moral priority of our

time.[11] Longtermism is about taking seriously just how big the future could be and how high the stakes are in shaping it. If humanity survives to even a fraction of its potential life span, then, strange as it may seem, we are the ancients: we live at the very beginning of history, in the most distant past. What we do now will affect untold numbers of future people. We need to act wisely.

It took me a long time to come around to longtermism. It's hard for an abstract ideal, focused on generations of people whom we will never meet, to motivate us as more salient problems do. In high school, I worked for organisations that took care of the elderly and disabled. As an undergraduate who was concerned about global poverty, I volunteered at a children's polio rehabilitation centre in Ethiopia. When starting graduate work, I tried to figure out how people could help one another more effectively. I committed to donating at least 10 percent of my income to charity, and I cofounded an organization, Giving What We Can, to encourage others to do the same.[12]

These activities had a tangible impact. By contrast, the thought of trying to improve the lives of unknown future people initially left me cold. When a colleague presented me with arguments for taking the long term seriously, my immediate reaction was glib dismissal. There are real problems in the world facing real people, I thought, problems like extreme poverty, lack of education, and death from easily preventable diseases. That's where we should focus. Sci-fi-seeming speculations about what might or might not impact the future seemed like a distraction.

But the arguments for longtermism exerted a persistent force on my mind. These arguments were based on simple ideas: that, impartially considered, future people should count for no less, morally, than the present generation; that there may be a huge number of future people; that life, for them, could be extraordinarily good or inordinately bad; and that we really can make a difference to the world they inhabit.

The most important sticking point for me was practical: Even if we should care about the longterm future, what can we do? But as I learned more about the potentially history-shaping

events that could occur in the near future, I took more seriously the idea that we might soon be approaching a critical juncture in the human story. Technological development is creating new threats and opportunities for humanity, putting the lives of future generations on the line.

I now believe the world's long-run fate depends in part on the choices we make in our lifetimes. The future could be wonderful: we could create a flourishing and long-lasting society, where everyone's lives are better than the very best lives today. Or the future could be terrible, falling to authoritarians who use surveillance and AI to lock in their ideology for all time, or even to AI systems that seek to gain power rather than promote a thriving society. Or there could be no future at all: we could kill ourselves off with biological weapons or wage an all-out nuclear war that causes civilisation to collapse and never recover.

There are things we can do to steer the future onto a better course. We can increase the chance of a wonderful future by improving the values that guide society and by carefully navigating the development of AI. We can ensure we get a future at all by preventing the creation or use of new weapons of mass destruction and by maintaining peace between the world's great powers. These are challenging issues, but what we do about them makes a real difference.

So I shifted my priorities. Still unsure about the foundations and implications of longtermism, I switched my research focus and cofounded two organisations to investigate these issues further: the Global Priorities Institute at Oxford University, and the Forethought Foundation. Drawing on what I have learned, I have tried to write the case for longtermism that would have convinced me a decade ago.

To illustrate the claims in this book, I rely on three primary metaphors throughout. The first is of humanity as an imprudent teenager. Most of a teenager's life is still ahead of them, and their decisions can have lifelong impacts. In choosing how much to study, what career to pursue, or which risks are too risky, they should think not just about short-term thrills but also about the whole course of the life ahead of them.

The second is of history as molten glass. At present, society is still malleable and can be blown into many shapes. But at some point, the glass might cool, set, and become much harder to change. The resulting shape could be beautiful or deformed, or the glass could shatter altogether, depending on what happens while the glass is still hot.

The third metaphor is of the path towards longterm impact as a risky expedition into uncharted terrain. In trying to make the future better, we don't know exactly what threats we will face or even exactly where we are trying to go; but, nonetheless, we can prepare ourselves. We can scout out the landscape ahead of us, ensure the expedition is well resourced and well coordinated, and, despite uncertainty, guard against those threats we are aware of.

This book's scope is broad. Not only am I arguing for longtermism; I'm also trying to work out its implications. I've therefore relied heavily on an extensive team of consultants and research assistants. Whenever I've stepped outside of moral philosophy, my area of expertise, domain experts have advised me from start to end. This book is therefore not really "mine": it has been a team effort. In total, this book represents over a decade's worth of full-time work, almost two years of which was spent fact-checking.

For those who want to dig deeper into some of my claims, I have compiled extensive supplementary materials, including special reports I commissioned as background research, and made them available at whatweowethefuture.com. Despite the work done so far, I believe we have only scratched the surface of longtermism and its implications; there is much still to learn.

If I'm right, then we face a huge responsibility. Relative to everyone who could come after us, we are a tiny minority. Yet we hold the entire future in our hands. Everyday ethics rarely grapples with such a scale. We need to build a moral worldview that takes seriously what's at stake.

By choosing wisely, we can be pivotal in putting humanity on the right course. And if we do, our great-great-grandchildren will look back and thank us, knowing that we did everything we could to give them a world that is just and beautiful.

CHAPTER 1

The Case for Longtermism

The Silent Billions

Future people count. There could be a lot of them. We can make their lives go better.

This is the case for longtermism in a nutshell. The premises are simple, and I don't think they're particularly controversial. Yet taking them seriously amounts to a moral revolution—one with far-reaching implications for how activists, researchers, policy makers, and indeed all of us should think and act.

Future people count, but we rarely count them. They cannot vote or lobby or run for public office, so politicians have scant incentive to think about them. They can't bargain or trade with us, so they have little representation in the market. And they can't make their views heard directly: they can't tweet, or write articles in newspapers, or march in the streets. They are utterly disenfranchised.

Previous social movements, such as those for civil rights and women's suffrage, have often sought to give greater recognition and influence to disempowered members of society. I see longtermism as an extension of these ideals. Though we cannot give genuine political power to future people, we can at least give consideration to them. By abandoning the tyranny of the present over the future, we can act as trustees—helping to create a flourishing world for generations to come. This is of the utmost importance. Let me explain why.

Future People Count

The idea that future people count is common sense. Future people, after all, are people. They will exist. They will have

hopes and joys and pains and regrets, just like the rest of us. They just don't exist *yet*.

To see how intuitive this is, suppose that, while hiking, I drop a glass bottle on the trail and it shatters. And suppose that if I don't clean it up, later a child will cut herself badly on the shards.[1] In deciding whether to clean it up, does it matter *when* the child will cut herself? Should I care whether it's a week, or a decade, or a century from now? No. Harm is harm, whenever it occurs.

Or suppose that a plague is going to infect a town and kill thousands. You can stop it. Before acting, do you need to know when the outbreak will occur? Does that matter, just on its own? No. The pain and death at stake are worthy of concern regardless.

The same holds for good things. Think of something you love in your own life; maybe it's music or sports. And now imagine someone else who loves something in their life just as much. Does the value of their joy disappear if they live in the future? Suppose you can give them tickets to see their favourite band or the football team they support. To decide whether to give them, do you need to know the delivery date?

Imagine what future people would think, looking back at us debating such questions. They would see some of us arguing that future people don't matter. But they look down at their hands; they look around at their lives. What is different? What is less real? Which side of the debate will seem more clear-headed and obvious? Which more myopic and parochial?

Distance in time is like distance in space. People matter even if they live thousands of miles away. Likewise, they matter even if they live thousands of years hence. In both cases, it's easy to mistake distance for unreality, to treat the limits of what we can see as the limits of the world. But just as the world does not stop at our doorstep or our country's borders, neither does it stop with our generation, or the next.

These ideas are common sense. A popular proverb says, "A society grows great when old men plant trees under whose shade they will never sit."[2] When we dispose of radioactive waste, we don't say, "Who cares if this poisons people centuries from now?"

Similarly, few of us who care about climate change or pollution do so solely for the sake of people alive today. We build museums and parks and bridges that we hope will last for generations; we invest in schools and longterm scientific projects; we preserve paintings, traditions, languages; we protect beautiful places. In many cases, we don't draw clear lines between our concerns for the present and the future—both are in play.

Concern for future generations is common sense across diverse intellectual traditions. The *Gayanashagowa*, the centuries-old oral constitution of the Iroquois Confederacy, has a particularly clear statement. It exhorts the Lords of the Confederacy to "have always in view not only the present but also the coming generations."[3] Oren Lyons, a faithkeeper for the Onondaga and Seneca nations of the Iroquois Confederacy, phrases this in terms of a "seventh-generation" principle, saying, "We... make every decision that we make relate to the welfare and well-being of the seventh generation to come.... We consider: will this be to the benefit of the seventh generation?"[4]

However, even if you grant that future people count, there's still a question of how much weight to give their interests. Are there reasons to care more about people alive today?

Two reasons stand out to me. The first is partiality. We often have stronger special relationships with people in the present, like family, friends, and fellow citizens, than with people in the future. It's common sense that you can and should give extra weight to your near and dear.

The second reason is reciprocity. Unless you live as a recluse in the wilderness, the actions of an enormous number of people—teachers, shopkeepers, engineers, and indeed all taxpayers—directly benefit you and have done so throughout your life. We typically think that if someone has benefited you, that gives you a reason to repay them. But future people don't benefit you the way others in your generation do.[5]

Special relationships and reciprocity are important. But they do not change the upshot of my argument. I'm not claiming that the interests of present and future people should always and everywhere be given equal weight. I'm just claiming that future people matter significantly. Just as caring more about our

children doesn't mean ignoring the interests of strangers, caring more about our contemporaries doesn't mean ignoring the interests of our descendants.

To illustrate, suppose that one day we discover Atlantis, a vast civilisation at the bottom of the sea. We realise that many of our activities affect Atlantis. When we dump waste into the oceans, we poison its citizens; when a ship sinks, they recycle it for scrap metal and other parts. We would have no special relationships with the Atlanteans, nor would we owe them repayment for benefits they had bestowed on us. But we should still give serious consideration to how our actions affect them.

The future is like Atlantis. It, too, is a vast, undiscovered country;[6] and whether that country thrives or falters depends, in significant part, on what we do today.

The Future Is Big

It's common sense that future people count. So, too, is the idea that, morally, the numbers matter. If you can save one person or ten from dying in a fire, then, all else being equal, you should save ten; if you can cure a hundred people or a thousand of a disease, you should cure a thousand. This matters, because the number of future people could be huge.

To see this, consider the long-run history of humanity. There have been members of the genus *Homo* on Earth for over 2.5 million years.[7] Our species, *Homo sapiens*, evolved around three hundred thousand years ago. Agriculture started just twelve thousand years ago, the first cities formed only six thousand years ago, the industrial era began around 250 years ago, and all the changes that have happened since then—transitioning from horse-drawn carts to space travel, leeches to heart transplants, mechanical calculators to supercomputers—occurred over the course of just three human lifetimes.[8]

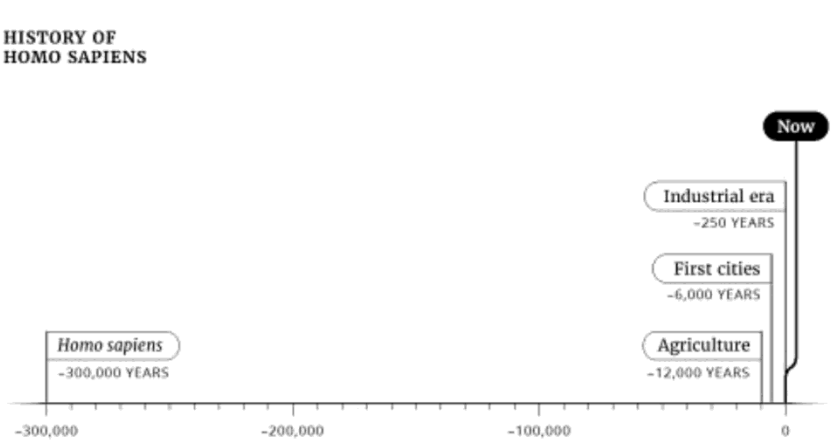


Figure 1.1. The history of Homo sapiens.

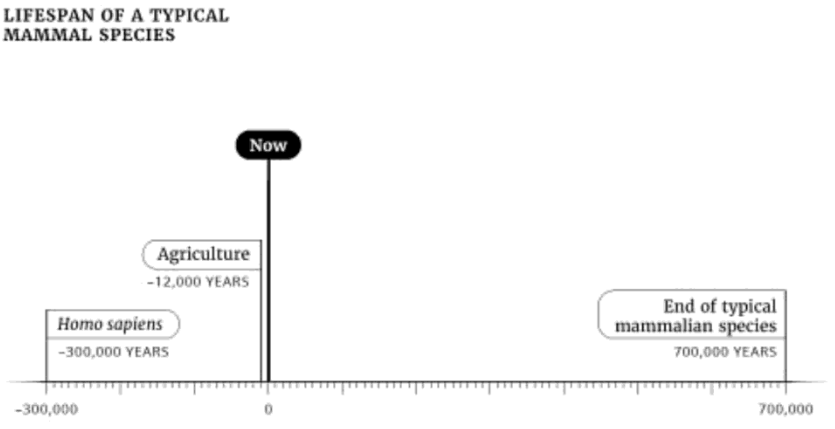


Figure 1.2. The potential future of civilisation, if humans survive as long as the average mammalian species.

How long will our species last? Of course, we don't know. But we can make informative estimates that take our uncertainty into account, including our uncertainty about whether we'll cause our own demise.

To illustrate the potential scale of the future, suppose that we only last as long as the typical mammalian species—that is, around one million years.[9] Also assume that our population continues at its current size. In that case, there would be eighty trillion people yet to come; future people would outnumber us ten thousand to one.

Of course, we must consider the whole range of ways the future could go. Our life span as a species could be much shorter than that of other mammals if we cause our own extinction. But it could also be much longer. Unlike other mammals, we have sophisticated tools that help us adapt to varied environments; abstract reasoning, which allows us to make complex, long-term plans in response to novel circumstances; and a shared culture that allows us to function in groups of millions. These help us avoid threats of extinction that other mammals can't.[10]

This has an asymmetric impact on humanity's life expectancy. The future of civilisation could be very short, ending within a few centuries. But it could also be extremely long. The earth will remain habitable for hundreds of millions of years. If we survive that long, with the same population per century as now, there will be a million future people for every person alive today. And if humanity ultimately takes to the stars, the timescales become literally astronomical. The sun will keep burning for five billion years; the last conventional star formations will occur in over a trillion years; and, due to a small but steady stream of collisions between brown dwarfs, a few stars will still shine a million trillion years from now.[11]

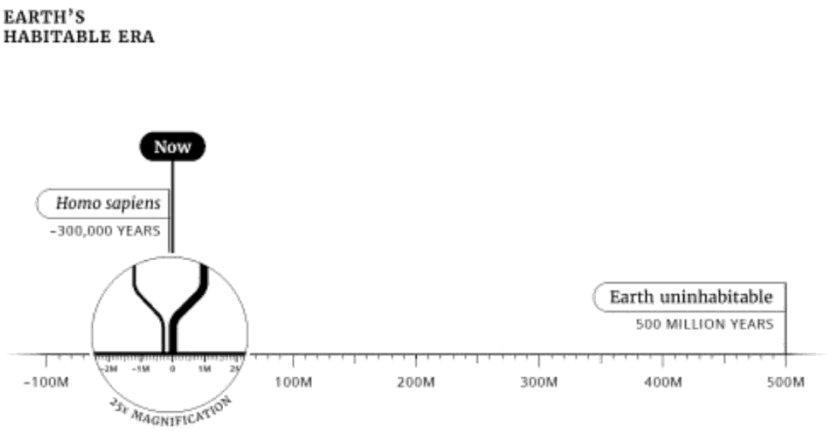


Figure 1.3. The potential future of civilisation if it survives until the earth becomes uninhabitable for humans due to the sun's increasing brightness. There is considerable uncertainty as to the length of this window, with estimates ranging from 500 million to 1.3 billion years.

The real possibility that civilisation will last such a long time gives humanity an enormous life expectancy. A 10 percent chance of surviving five hundred million years until the earth is no longer habitable gives us a life expectancy of over fifty million years; a 1 percent chance of surviving until the last conventional star formations give us a life expectancy of over ten billion years.[12]

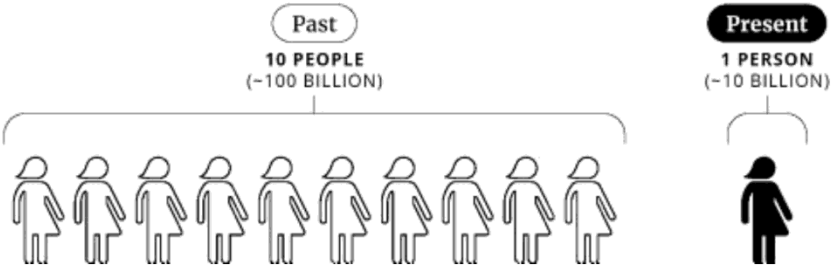
Ultimately, we shouldn't care just about humanity's life expectancy but also about how many people there will be. So we must ask: How many people in the future will be alive at any one time?

Future populations might be much smaller or much larger than they are today. But if the future population is smaller, it can be smaller by eight billion at most—the size of today's population. In contrast, if the future population is bigger, it could be much bigger. The current global population is already over a thousand times larger than it was in the hunter-gatherer era. If global population density increased to that of the Netherlands—an agricultural net exporter—there would be seventy billion people alive at any one time.[13] This might seem fantastical, but a global population of eight billion would have seemed fantastical to a prehistoric hunter-gatherer or an early agriculturalist.

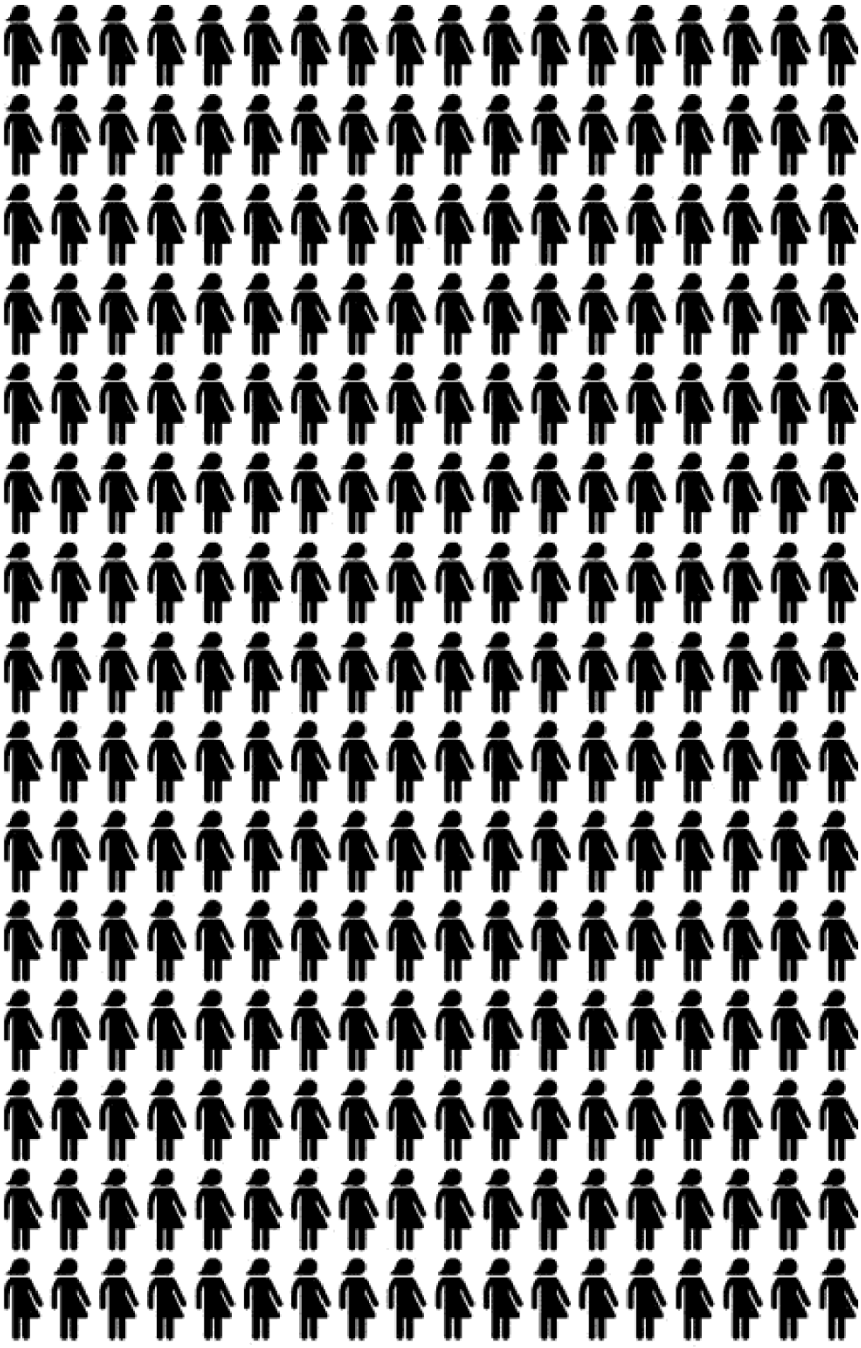
Population size could get dramatically larger again if we one day take to the stars. Our sun produces billions of times as much sunlight as lands on Earth, there are tens of billions of other stars across our galaxy, and billions of galaxies are accessible to us.[14] There might therefore be vastly more people in the distant future than there are today.

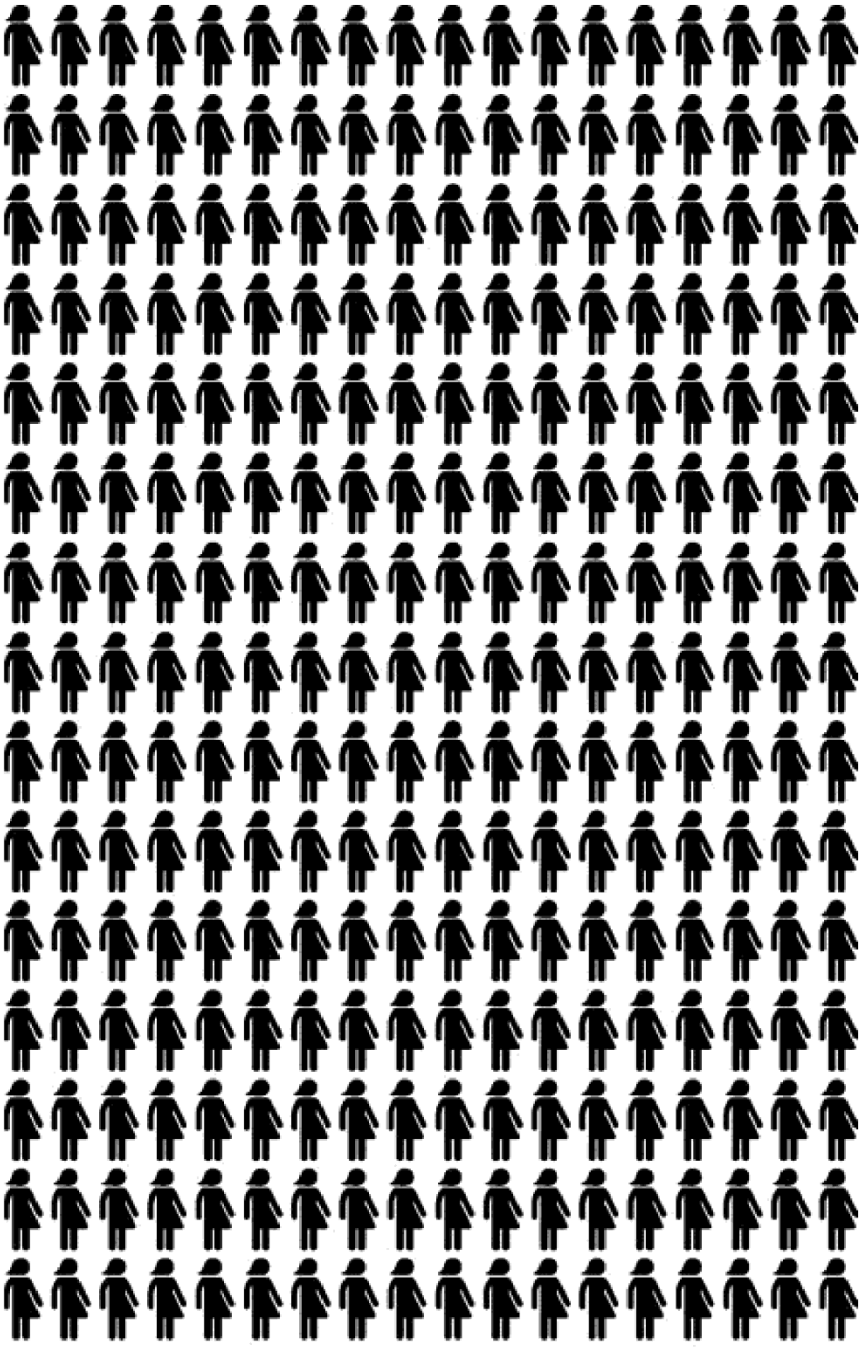
Just how many? Precise estimates are neither possible nor necessary. On any reasonable accounting, the number is immense.

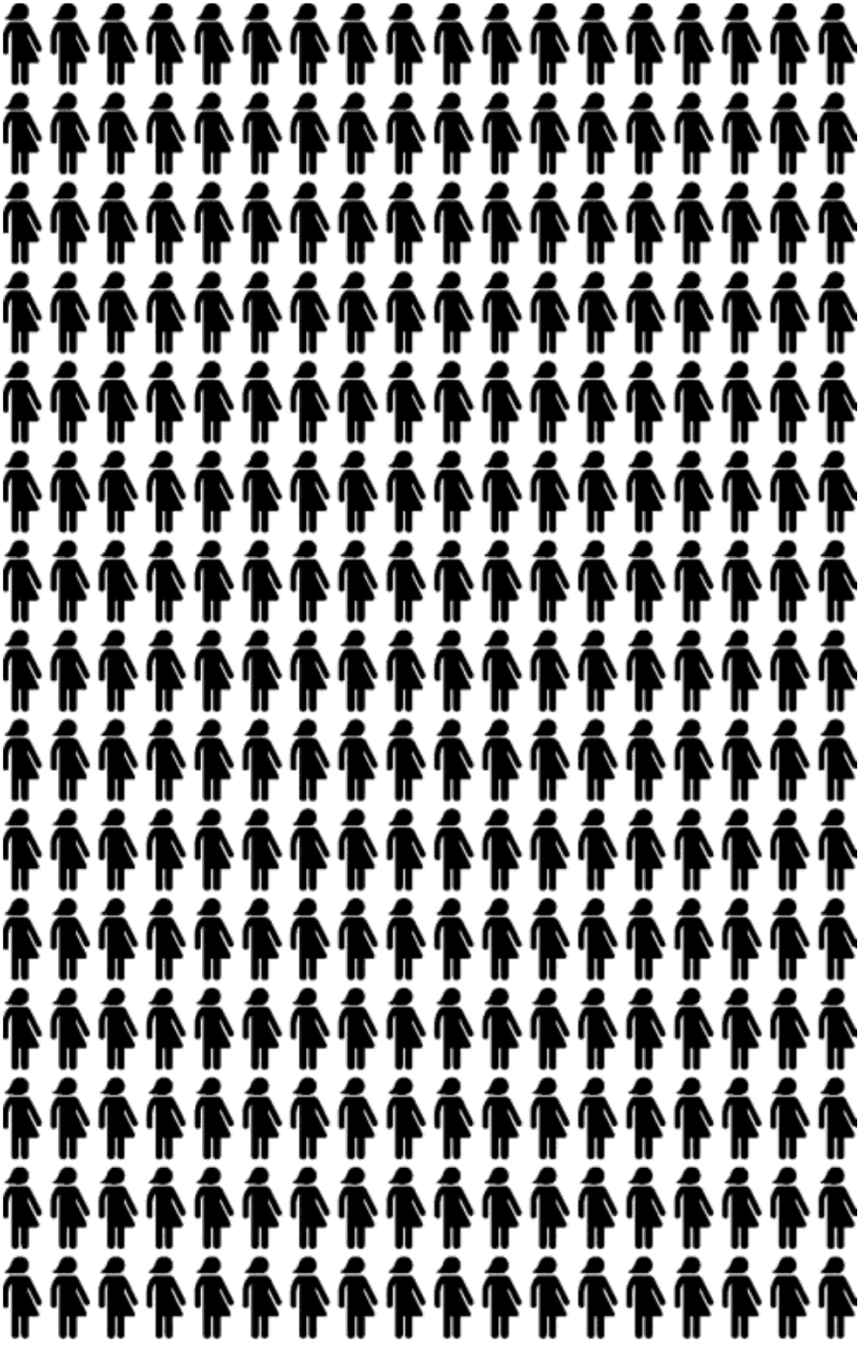
To see this, look at the following diagram. Each figure represents ten billion people. So far, roughly one hundred billion people have ever lived. These past people are represented as ten figures. The present generation consists of almost eight billion people, which I'll round up to ten billion and represent with a single figure:



Next, we'll represent the future. Let's just consider the scenario where we stay at current population levels and live on Earth for five hundred million years. These are all the future people:







Represented visually, we begin to see how many lives are at stake. But I cut the diagram short. The full version would fill twenty thousand pages—saturating this book a hundred times over. Each figure would represent ten billion lives, and each of those lives could be flourishing or wretched.

Earlier, I suggested that humanity today is like an imprudent teenager: most of our life is ahead of us, and decisions that impact the rest of that life are of colossal importance. But, really, this analogy understates my case. A teenager knows approximately how long she can expect to live. But we do not know humanity's life expectancy. We are more like a teenager who, for all she knows, might accidentally cause her own death in the next few months but also might live for a thousand years. If you were in such a situation, would you think seriously about the long life that might be ahead of you, or would you ignore it?

The sheer size of the future can be dizzying. Typically, "longterm" thinking involves attention to years or decades at most. But even with a low estimate of humanity's life expectancy, this is like a teenager believing that longterm thinking means considering tomorrow but not the day after.

Despite how overwhelming thoughts of our future can be, if we truly care about the interests of future generations—if we recognize that they are real people, capable of happiness and suffering just like us—then we have a duty to consider how we might impact the world they inhabit.

The Value of the Future

The future could be very big. It could also be very good—or very bad.

To get a sense of how good, we can look at some of the progress humanity has made over the last few centuries. Two hundred years ago, average life expectancy was less than thirty; today, it is seventy-three.[15] Back then, over 80 percent of the world lived in extreme poverty; now, less than 10 percent does.[16] Back then, only about 10 percent of adults could read; today, more than 85 percent can.[17]

Collectively we have the power both to encourage these positive trends and to change course on the negative trends, like the dramatic increases in carbon dioxide emissions and in the number of animals suffering in factory farms. We can build a world where everyone lives like the happiest people in the most well-off countries today, a world where no one lives in poverty, no one lacks adequate medical care, and, insofar as is possible, everyone is free to live as they want.

But we could do even better still—far better. The best that we have seen so far is a poor guide to what is possible. To get some inkling of this, consider the life of a rich man in Britain in 1700—a man with access to the best food, health care, and luxuries available at the time. For all his advantages, such a man could easily die of smallpox, syphilis, or typhus. If he needed surgery or had a toothache, the treatment would be agonising and carry a significant risk of infection. If he lived in London, the air he breathed would be seventeen times as polluted as it is today.[18] Travelling even within Britain could take weeks, and most of the globe was entirely inaccessible to him. If he had imagined a future merely where most people were as rich as him, he would have failed to anticipate many of the things that improve our lives, like electricity, anaesthesia, antibiotics, and modern travel.

It's not just technology that has improved people's lives; moral change has done so, too. In 1700, women were unable to attend university, and the feminist movement did not exist.[19] If that well-off Brit was gay, he could not love openly; sodomy was punishable by death.[20] In the late 1700s, three in four people

globally were the victims of some form of forced labour; now less than 1 percent are.[21] In 1700, no one lived in a democracy. Now over half the world does.[22]

Much of the progress we've made since 1700 would have been very difficult for people back then to anticipate. And that's with only a three-century gap. Humanity could last for millions of centuries on Earth alone. On such a scale, if we anchor our sense of humanity's potential to a fixed-up version of our present world, we risk dramatically underestimating just how good life in the future could be.

Consider the very best moments in your life—moments of joy, beauty, and energy, like falling in love, or achieving a lifelong goal, or having some creative insight. These moments provide proof of what is possible: we know that life can be at least as good as it is then. But they also show us a direction in which our lives can move, leading somewhere we have yet to go. If my best days can be hundreds of times better than my typically pleasant but humdrum life, then perhaps the best days of those in the future can be hundreds of times better again.

I'm not claiming that a wonderful future is *likely*. Etymologically, "utopia" means "no-place," and indeed the path from here to some ideal future state is very fragile. But a wonderful future is not just a fantasy, either. A better word would be "eutopia," meaning "good place"—something to strive for. It's a future that, with enough patience and wisdom, our descendants could actually build—if we pave the way for them.

And though the future could be wonderful, it could also be terrible. To see this, look at some of the negative trends of the past and imagine a future where *they* are the dominant forces guiding the world. Consider that slavery had all but disappeared from France and England by the end of the twelfth century, but in the colonial era those same countries became slave traders on a massive scale.[23] Or consider that the mid-twentieth century saw totalitarian regimes emerging even out of democracies. Or that we used scientific advances to build nuclear weapons and factory farms.

Just as eutopia is a real possibility, so is dystopia. The future could be one where a single totalitarian regime controls the

world, or where today's quality of life is but a distant memory of a former Golden Age, or where a third world war has led to the complete destruction of civilisation. Whether the future is wonderful or terrible is, in part, up to us.

Not Just Climate Change

Even if you accept that the future is big and important, you might be skeptical that we can positively affect it. And I agree that working out the long-run effects of our actions is very hard. There are many considerations at play, and our understanding of them is just beginning. My aim with this book is to stimulate further work in this area, not to be definitive in any conclusions about what we should do. But the future is so important that we've got to at least try to figure out how to steer it in a positive direction. And, already, there are some things we can say.

Looking to the past, though there are not many examples of people deliberately aiming at long-run impacts, they do exist, and some had surprising levels of success. Poets provide one source. In Shakespeare's Sonnet 18 ("Shall I compare thee to a summer's day?") the author notes that through his art he can preserve the young man he admires for all eternity:[24]

> *But thy eternal summer shall not fade,*
> .....
> *When in eternal lines to time thou grow'st.*
> *So long as men can breathe or eyes can see,*
> *So long lives this, and this gives life to thee.*[25]

Sonnet 18 was written in the 1590s but echoes a tradition that goes back much further.[26] In 23 BC the Roman poet Horace began the final poem in his *Odes* with these lines:[27]

> I have finished a monument more lasting than bronze, more lofty than the regal structure of the pyramids, one which neither corroding rain nor the ungovernable

> North Wind can ever destroy, nor the countless series of the years, nor the flight of time.
>
> I shall not wholly die, and a large part of me will elude the Goddess of Death.[28]

These claims seem bombastic, to say the least. But, plausibly, these poets' attempts at immortality succeeded. They have survived many hundreds of years and are in fact flourishing as the years pass: more people read Shakespeare today than did in his own time, and the same is probably true of Horace. And as long as some member of each future generation is willing to pay the tiny cost involved in preserving or replicating some representation of these poems, they will persist forever.

Other writers have also successfully aimed at very longterm impact. Thucydides wrote his *History of the Peloponnesian War* in the fifth century BC.[29] Many consider him the first Western historian to try to depict events faithfully and analyse their causes.[30] He believed he was describing general truths, and he deliberately wrote his history so that it could be influential far into the future:

> It will be enough for me, however, if these words of mine are judged useful by those who want to understand clearly the events which happened in the past and which (human nature being what it is) will, at some time or other and in much the same ways, be repeated in the future. My work is not a piece of writing designed to meet the taste of an immediate public, but was done to last for ever.[31]

Thucydides's work is still enormously influential to this day. It is required reading at the West Point and Annapolis military academies and the US Naval War College.[32] The widely read 2017 book *Destined for War*, by political scientist Graham Allison, had the subtitle *Can America and China Escape Thucydides's Trap?* Allison analyses US-China relations in the same terms that Thucydides used for Sparta and Athens. As far as I know,

Thucydides is the first person in recorded history to have deliberately aimed at longterm impact and succeeded.

More recent examples come from the United States' Founding Fathers. The US Constitution is almost 250 years old and has mostly remained the same throughout its life. Its founding was of enormous longterm importance, and many of the Founding Fathers were well aware of this. John Adams, the second president of the United States, commented, "The institutions now made in America will not wholly wear out for thousands of years. It is of the last importance, then, that they should begin right. If they set out wrong, they will never be able to return, unless it be by accident, to the right path."[33]

Similarly, Benjamin Franklin had such a reputation for believing in the health and longevity of the United States that in 1784 a French mathematician wrote a friendly satire of him, suggesting that if Franklin was sincere in his beliefs, he should invest his money to pay out on social projects centuries later, getting the benefits of compound interest along the way.[34] Franklin thought it was a great idea, and in 1790 he invested £1000 (about $135,000 in today's money) each for the cities of Boston and Philadelphia: three-quarters of the funds would be paid out after one hundred years, and the remainder after two hundred years. By 1990, when the final funds were distributed, the donation had grown to almost $5 million for Boston and $2.3 million for Philadelphia.[35]

The Founding Fathers themselves were influenced by ideas developed almost two thousand years before them. Their views on the separation of powers were foreshadowed by Locke and Montesquieu, who drew on Polybius's analysis of Roman governance from the second century BC.[36] We also know that several Founding Fathers were familiar with Polybius's work themselves.[37]

Those of us in the present don't need to be as influential as Thucydides or Franklin to predictably impact the longterm future. In fact, we do it all the time. We drive. We fly. We thereby emit greenhouse gases with very long-lasting effects. Natural processes will return carbon dioxide concentrations to

preindustrial levels only after hundreds of thousands of years.[38] These are timescales usually associated with radioactive nuclear waste.[39] However, with nuclear power we carefully store and plan to bury the waste products; with fossil fuels we belch them into the air.[40]

In some cases, the geophysical impacts of this warming get even more extreme over time rather than "washing out."[41] The Intergovernmental Panel on Climate Change (IPCC) projects that in the medium-low-emissions scenario, which is now widely seen to be the most likely, sea level would rise by around 0.75 metres by the end of the century.[42] But it would keep rising well past the year 2100. After ten thousand years, sea level would be ten to twenty metres higher than it is today.[43] Hanoi, Shanghai, Kolkata, Tokyo, and New York would all be mostly below sea level.[44]

Climate change shows how actions today can have longterm consequences. But it also highlights that longterm-oriented actions needn't involve ignoring the interests of those alive today. We can positively steer the future while improving the present, too.

Moving to clean energy has enormous benefits in terms of present-day human health. Burning fossil fuels pollutes the air with small particles that cause lung cancer, heart disease, and respiratory infections.[45] As a result, every year about 3.6 million people die prematurely.[46] Even in the European Union, which in global terms is comparatively unpolluted, air pollution from fossil fuels causes the average citizen to lose a whole year of life.[47]

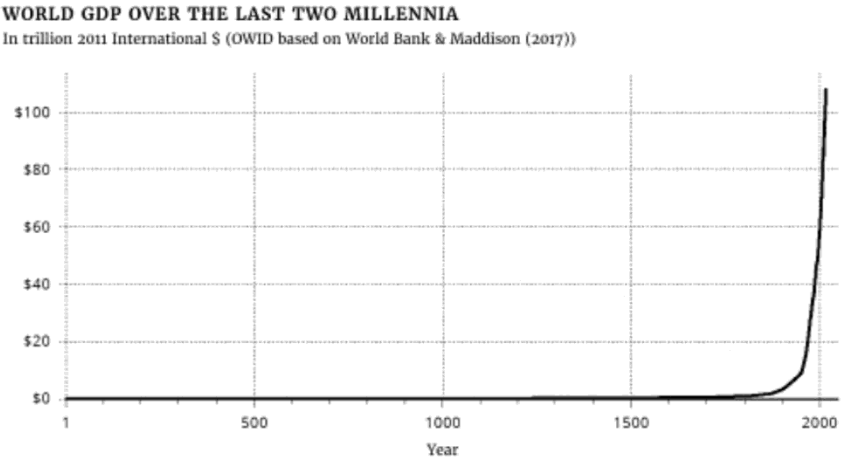


Figure 1.5. World economic output since AD 1.

So we know that the present era is extremely unusual compared to the past. But it's also unusual compared to the future. This rapid rate of change cannot continue forever, even if we entirely decouple growth from carbon emissions and even if in the future we spread to the stars. To see this, suppose that future growth slows a little to just 2 percent per year.[54] At such a rate, in ten thousand years the world economy would be 10^{86} times larger than it is today—that is, we would produce one hundred trillion trillion trillion trillion trillion trillion trillion times as much output as we do now. But there are less than 10^{67} atoms within ten thousand light years of Earth.[55] So if current growth rates continued for just ten millennia more, there would have to be ten million trillion times as much output as our current world produces for *every atom* that we could, in principle, access. Though of course we can't be certain, this just doesn't seem possible.[56]

Humanity might last for millions or even billions of years to come. But the rate of change of the modern world can only continue for thousands of years. What this means is that we are living through an extraordinary chapter in humanity's story. Compared to both the past and the future, every decade we live through sees an extremely unusual number of economic and technological changes. And some of these changes—like the

inventions of fossil fuel power, nuclear weapons, engineered pathogens, and advanced artificial intelligence—have the potential to impact the whole course of the future.

It's not only the rapid rate of change that makes this time unusual. We're also unusually connected.[57] For over fifty thousand years, we were broken up into distinct groups; there was simply no way for people across Africa, Europe, Asia, or Australia to communicate with one another.[58] Between 100 BC and AD 150 the Roman Empire and the Han dynasty each comprised up to 30 percent of the world's population, yet they barely knew of each other.[59] Even within one empire, one person had very limited ability to communicate with someone far away.

In the future, if we spread to the stars, we will again be separated. The galaxy is like an archipelago, vast expanses of emptiness dotted with tiny pinpricks of warmth. If the Milky Way were the size of Earth, our solar system would be ten centimetres across and hundreds of metres would separate us from our neighbours. Between one end of the galaxy and the other, the fastest possible communication would take a hundred thousand years; even between us and our closest neighbour, there-and-back communication would take almost nine years.[60]

In fact, if humanity spreads far enough and survives long enough, it will eventually become impossible for one part of civilisation to communicate with another. The universe is composed of millions of groups of galaxies.[61] Our own is called, simply, the Local Group. The galaxies within each group are close enough to each other that gravity binds them together forever.[62] But, because the universe is expanding, the groups of galaxies will eventually be torn apart from each other. Over 150 billion years in the future, not even light will be able to travel from one group to another.[63]

The fact that our time is so unusual gives us an outsized opportunity to make a difference. Few people who ever live will have as much power to positively influence the future as we do. Such rapid technological, social, and environmental change means that we have more opportunity to affect when and how the most important of these changes occur, including by

managing technologies that could lock in bad values or imperil our survival. Civilisation's current unification means that small groups have the power to influence the whole of it. New ideas are not confined to a single continent, and they can spread around the world in minutes rather than centuries.

The fact that these changes are so recent means, moreover, that we are out of equilibrium: society has not yet settled down into a stable state, and we are able to influence *which* stable state we end up in. Imagine a giant ball rolling rapidly over a rugged landscape. Over time it will lose momentum and slow, settling at the bottom of some valley or chasm. Civilisation is like this ball: while still in motion, a small push can affect in which direction we roll and where we come to rest.

CHAPTER 2

You Can Shape the Course of History

Prehistory's Impact on Today

Human beings have been making choices with longterm consequences for tens of thousands of years. Consider: Why is Africa home to so many more species of megafauna—large animals like elephants and giraffes—than the rest of the world?[1] You might think, as I did before learning about this topic, that the answer has to do with Africa's particular environment. But that's not right. Fifty thousand years ago, a great variety of megafauna roamed the planet.

Consider the glyptodonts, a group of armadillo-like herbivores that lived in South America for tens of millions of years.[2] The largest glyptodonts were as big and heavy as cars.[3] Their bodies were encased in a giant shell, they had a bone helmet, and some of them had club-shaped tails adorned with spikes.[4] They looked like giant capybaras dressed up as armoured trucks. They went extinct around 12,000 years ago.[5]

Or consider megatherium, a giant ground sloth and one of the largest land mammals to have ever lived, rivalling the Asian elephant in size.[6] It went extinct 12,500 years ago.[7] Or *Notiomastodon*, a genus of elephant-like animals with giant tusks that evolved two million years ago and went extinct 10,000 years ago.[8] Or the dire wolf, the largest known canine to have lived, which, having lost its giant herbivorous prey, went extinct 13,000 years ago.[9] All these species lived in South America, along with dozens of other megafauna species that are no longer with us.

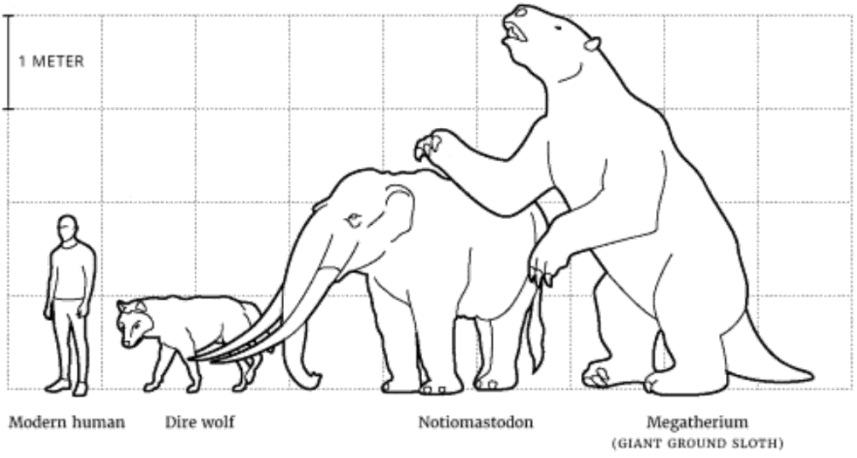


Figure 2.1. Some specimens of now-extinct megafauna drawn at scale in comparison with a modern human.

There is a heated debate over what caused the extinctions of megafauna. Some scientists believe that natural climate change was the main driver, some believe that humans were the culprit, and some believe it was a mix of humans and climate change.[10] In my view, the evidence is clear that humans often played a decisive role: most of these megafauna survived over a dozen similarly sized climatic changes in the past;[11] smaller animals did not go extinct at nearly the same rate as megafauna;[12] and the timing of their extinction usually coincides with humans' arrival into their habitats.[13] Though perhaps helped by climate change, it was hunting and the disruption of natural environments caused by human activity that killed them off. Unlike megafauna on other continents, African megafauna evolved alongside humans and so were better prepared for *Homo sapiens* as a predator.

The extinction of these megafauna was probably an irrevocable change to the world, made by humans with extremely primitive technology. It meant we lost, for all time, many beautiful and unique species. And *Homo sapiens* are not only implicated in the extinction of giant sloths and canines: we are also the prime suspect in the end of our human cousins, the Denisovans and the Neanderthals, who likely died out as a result

else. Agriculture is less contingent because it emerged in multiple locations independently.

If something is very noncontingent, then the change would have happened soon anyway, even without the individual's action. Knowledge of calculus was not very contingent because Leibniz independently discovered it just a few years after Newton did. Considering contingency is crucial because if you make a change to the world but it's a change that would have simply happened soon afterward anyway, then you have not made a longterm *difference* to the world.

Though it's hard to be confident, my guess is that the extinction of the glyptodonts was not very contingent. Even if the hunters who killed off the last of them had not done so, then probably some other group of hunters, at some later time, would have. In order to prevent the glyptodonts' extinction, those hunters would have had to promote a norm that the glyptodonts should be protected and this norm would have had to be passed down the generations, and adhered to, until the present day. This would not be impossible to pull off, but it does seem difficult.

Multiplying significance, persistence, and contingency together gives us the longterm value of bringing about some state of affairs. Because of this, we can make intuitive comparisons between different longterm effects on these dimensions. For example, between two alternatives, if one is ten times as persistent as the other, that will outweigh the other being eight times as significant. Because the potential scale of the longterm future is so great—millions, billions, or even trillions of years—our attention should be, first, on what states of affairs might be the most persistent. Then, afterwards, we can think about significance and contingency.

Table 2.1. The Significance, Persistence, Contingency Framework

Significance	What's the average value added by bringing about a certain state of affairs?
Persistence	How long will this state of affairs last once it has been brought about?

Contingency	If not for the action under consideration, how briefly would the world have been in this state of affairs (if ever)?

Note: For more details, see Appendix 3.

To see how this framework can be used to guide our decisions today, let's return to the metaphor of humanity as an imprudent teenager. Looking back at our own individual teenage years, what choices mattered most? Plausibly, it's those whose effects were the most persistent, affecting the whole course of our lives; most significant, making the biggest difference to our wellbeing at any one time; and most contingent, causing an effect that would not have happened anyway at some later date.

Some choices I made as a teenager did not have persistent effects: my plans for the weekend made a difference to that weekend but usually didn't shape the course of my life. The effects of other choices were not that contingent. Like many teenagers, I cared about firsts—first drink, first time having sex. But ultimately, such firsts would have happened at some point regardless, and looking back, the precise timing did not matter much. Finally, some effects, though persistent and contingent, just weren't that significant. I chose not to get braces to close the gap between my two front teeth because at the time I believed that a gap brings good luck. I still have the gap today, but as far as I can tell, it has not significantly affected my life.

Other decisions I made mattered a lot. I was reckless as a teenager and sometimes went "buildering," also known as urban climbing. Once, coming down from the roof of a hotel in Glasgow, I put my foot on a skylight and fell through. I caught myself at waist height, but the broken glass punctured my side. Luckily, it missed all internal organs. A little deeper, though, and my guts would have popped out violently, and I could easily have died. I still have the scar: three inches long and almost half an inch thick, curved like an earthworm. Dying that evening would have prevented all the rest of my life. My choice to go buildering was therefore an enormously important (and enormously foolish) decision—one of the highest-stakes decisions I'll ever make.

More mundanely, I could easily have exposed myself to a different set of intellectual influences, which would have set me

on a very different path in life. All my close friends studied medicine—the standard path for smart, socially minded teenagers in Scotland—and I considered it for myself. If I had not studied philosophy at school, and if I hadn't had such an engaged and passionate teacher, Jeremy Hall, I would probably not have studied it at university or pursued it as a career. I expect that a career in medicine would have been fulfilling, but it probably would not have exposed me to the moral arguments that led me to the path I've taken—a difference which, from my current perspective, would have been a major loss.

Looking back, it's clear that, for many of my teenage choices, what mattered most was not the fun I had at the time—whether buildering was a thrill (it was) or whether studying medicine at Edinburgh involved better parties. Rather, what mattered most was the impact of these choices on the rest of my life, whether I was risking death or altering the values that would guide my future self.

The risk of death I bore as a teenager and the intellectual influences that shaped my life mirror the two main ways in which we can impact the longterm future. First, we can affect humanity's duration: ensuring that we survive the next few centuries affects how many future generations there are. That is, we can help *ensure civilisation's survival.* Just as my teenage decisions to gamble with my life were among the most consequential I've ever made, so too are our decisions about how to handle risks of extinction or unrecovered civilisational collapse among the most consequential decisions that we as a society make today.

Second, we can affect civilisation's average value, changing how well or badly life goes for future generations, potentially for as long as civilisation lasts. That is, we can *change trajectory*, trying to improve the quality of future people's lives over the life span of civilisation.[20] Just as the intellectual influences I was exposed to as a teenager shaped the whole rest of my life, so, too, I will argue, the values that humanity adopts in the next few centuries might shape the entire trajectory of the future.[21]

These two ideas structure the book. Part II of this book looks at trajectory changes, focusing in particular on changing

society's values. Within this, Chapter 3 argues for the significance and contingency of value changes, focusing on the abolition of slavery as a case study. Chapter 4 argues for the persistence of values, suggesting that new technology, in particular advanced artificial intelligence, could enable those in power to lock in their values indefinitely. Whether the future is governed by values that are authoritarian or egalitarian, benevolent or sadistic, exploratory or rigid, might well be determined by what happens this century.

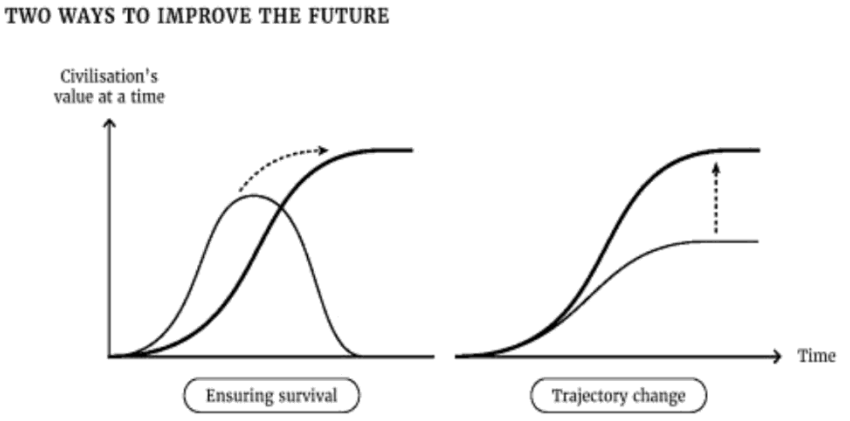


Figure 2.2. We can make the future better in two ways: by averting permanent catastrophes, thereby ensuring civilisation's survival; or by changing civilisation's trajectory to make it better while it lasts.

Part III looks at three ways of ensuring survival, dedicating a chapter to each. The first way is to prevent direct risks of human extinction; I focus on engineered pandemics. The second is to prevent the unrecovered collapse of civilisation; I focus on risks from nuclear war and extreme climate change. The third is technological stagnation, which could increase the risks of both extinction and collapse. Along the way, I discuss the persistence and contingency of the end of civilisation.

The question of the *significance* of the end of civilisation raises philosophical issues. Broadly, ensuring survival increases the quantity of future life; trajectory changes increase its quality. But you might not care much about sheer quantity. If there's no longer anyone around to care, why should it matter if civilisation has ended? And maybe, on balance, the future is more bad than good. If these worries were correct, then the

longtermist priority should be to increase the average value of future civilisation rather than its duration. Improving our trajectory would be more important than ensuring survival.

Part IV tackles these issues. I argue both that we should think of the nonexistence of future generations as a moral loss, if the people in them would have sufficiently good lives, and that we should expect the future to be more good than bad, on balance. Ensuring survival is therefore just as great a priority as improving our trajectory.

Part V turns to action. Longtermism is not just abstract philosophical speculation. It's an idea that people are putting into practice today. Chapter 10 looks at what some people are doing today to try to make the long term better, and how you can help.[22]

Thinking in Bets

When thinking about the changes that we could make to the world, we will not know how long they will last or how significant or contingent they will be. So we need a way of making decisions in the face of uncertainty. The most widely accepted account of how to do so is expected value theory.

Over the course of writing this book, I was repeatedly and viscerally reminded of the idea of expected value theory by my housemate at the time, Liv Boeree. Liv is one of the most successful female poker players of all time—a European Poker Tour and World Series champion. Her understanding and internalisation of the idea of expected value—or "EV," as she calls it—is critical to her success.

There are three aspects to expected value. First, probabilities. Rather than thinking that a three-of-a-kind poker hand is "very unlikely," Liv knows that the chance of getting one, before any cards are dealt, is about 5 percent; if the first two cards she's dealt are a pair, this probability rises to about 12 percent.[23] Though both probabilities are small, the difference between them can easily be enough to affect your decisions at the poker table.

Expected value theory is not just useful when gambling. It's crucial whenever we have to take a bet—that is, to make a decision in the face of uncertainty—which is almost all the time. My teenage decisions make this vivid. Before going buildering, I dismissed the possibility of falling and dying as unlikely and therefore not worth worrying about. But that was hugely foolish—not because it was *likely* that I would fall and die, but because it wasn't *sufficiently unlikely*, and dying is so bad that even a small chance is well worth avoiding.

In the face of an uncertain future, humanity often acts like my reckless teenage self. For example, climate change sceptics often point to our uncertainty as a reason for inaction.[26] There's so much we don't know, they claim—we don't know exactly how well climate models predict the amount of warming for a given quantity of emissions, for instance, or just how damaging a certain amount of warming would be for the economy. So we should not waste resources on the problem. But this is a terrible argument. We can grant that there's great uncertainty about what climate change means. But uncertainty cuts both ways. The damage caused by climate change might be less than is typically forecasted, but it might also be considerably *worse*—if, for example, the climate is more sensitive to temperature changes than such forecasts presuppose, or adaptation is harder, or we will emit more carbon dioxide than experts currently predict.

Crucially, the uncertainty around climate change is not symmetric: greater uncertainty should prompt more concern about worst-case outcomes, and this shift is not offset by a higher chance of best-case outcomes, because the worst-case outcomes are worse than the best-case outcomes are good.[27] For example, according to the Intergovernmental Panel on Climate Change, on the medium-low-emissions scenario, the best guess is that we will end up with around 2.5 degrees Celsius of warming by the end of the century.[28] But this is uncertain. There is a one-in-ten chance that we get 2 degrees or less. But that should not reassure us, because there is also a one-in-ten chance that we get more than 3.5 degrees.[29] Less than 2 degrees would be

something of a relief compared to the best-guess estimate, but more than 3.5 degrees would be much worse. The uncertainty gives us *more* reason to worry, not less. It's as if my teenage self, before jumping off a building, had reassured onlookers by saying, "It's OK, I've no idea how far I'll fall!"

Much the same will be true for the issues that I cover in this book. I'm not saying that we should be confident that value lock-in or major catastrophe will occur this century. What I am saying is that their chance of occurring is very real—certainly more than 1 percent, and certainly greater than many everyday risks, like dying in a car crash. When combined with how much is at stake, the expected value of trying to ensure a good future is enormous.

When we're applying the significance, persistence, and contingency framework, we should therefore be thinking about expected significance, expected persistence, and expected contingency.[30] If some change to the world has an 80 percent chance of fizzling out after ten years but a 20 percent chance of lasting for a million years, then its expected persistence is over two hundred thousand years. In general, if some change to the world has at least a reasonable chance of being highly significant, persistent, and contingent, then that can be sufficient for the expected value of that change to be very great indeed.

Moments of Plasticity

Often, some event can have highly significant, persistent, and contingent effects if there is a period of plasticity, where ideas or events or institutions can take one of many forms, followed by a period of rigidity or ossification. The dynamic is like that of glassblowing: In one period, the glass is still molten and malleable; it can be blown into one of many shapes. After it cools, it becomes rigid, and further change is impossible without remelting.

Plasticity frequently comes after a crisis, like a war. For example, after the end of World War II, Korea was divided along the thirty-eighth parallel. The location of the division was extremely contingent. Colonel Dean Rusk and Charles Bonesteel,

two American officers in their midthirties using a *National Geographic* map, proposed the thirty-eighth parallel because it divided the country roughly in half while keeping Seoul on the American side.[31] They were working on short notice because the United States had to reach an agreement with the Soviet Union before the entire peninsula fell into Soviet hands. No experts on Korea were consulted, and the proposed border cut across several preexisting Korean provinces and geographic features. In fact, the United States was surprised that the Soviets accepted the division; not only did it give Seoul to the United States, but Soviet troops were already in Korea while the closest American forces were still in Okinawa, several hundred miles away.[32] Yet after the division was implemented, it became hard to reverse, and it has since resulted in enormous differences to the fates of those who ended up in each of those two countries. South Koreans live in a strong democracy and are almost thirty times richer on average than they were in 1953. North Koreans live under a totalitarian dictatorship and may be even poorer than they were before the Korean War.[33]

A period of plasticity also commonly occurs when some idea or institution is still new. For example, the US Constitution was written over just four months—a moment of great plasticity—and amended eleven times in its first six years of operation.[34] After that, though, it became more rigid. Between 1804 and 1913, only three amendments were passed, all immediately following the Civil War: they abolished slavery, granted citizenship to African Americans and formerly enslaved people, and prohibited race from influencing the right to vote.[35] Today, the Constitution is again very rigid: it's only been amended once in the last fifty years, and that amendment—to prevent increases in congressional salaries from taking effect until the next term of office—was first proposed in 1789.[36]

This dynamic can hold for the laws and norms relevant to new technologies, too. Following World War II, the international community debated a variety of ways nuclear weapons could be governed.[37] One proposal, put forward by the United States, was the Baruch Plan, according to which the United States would

disband its nuclear weapons programme and transfer its bombs to the UN to be destroyed. The UN would then oversee the mining of fissionable materials around the world and inspect other countries to ensure that no one was building nuclear bombs. The USSR countered with the Gromyko Plan, which also proposed universal disarmament. Both of these plans failed, and it's not clear that either ever had much of a chance. But it was clearly a time of much greater plasticity in nuclear governance than we see now. Today, the idea that the UN could control the mining of uranium seems entirely off the table.

The dynamic of "early plasticity, later rigidity" can hold for new ideas, too. In addition to the books that we now know as the New Testament, a number of other texts were taught by some early Christians.[38] The New Testament books became the core Christian teachings only over the course of the first and second centuries AD and were not cemented until around the end of the fourth century AD.[39]

A final example comes from the history of climate change activism. The effect that carbon dioxide would have on global warming was first quantified in 1896 by Svante Arrhenius; his 1906 estimate of equilibrium climate sensitivity was four degrees, which is only a little higher than modern estimates.[40] And it was knowable, at that time, that we would probably emit dramatically more carbon dioxide in the future: one simply needed to continue extrapolating the trend of exponential economic growth and to recognize the obvious fact that such growth would bring a corresponding increase in energy demand.

In 1958, Frank Capra, director of *It's a Wonderful Life*, made an educational weather documentary, *Unchained Goddess*, which included a warning about climate change: "Even now, man may be unwittingly changing the world's climate through the waste products of his civilisation. Due to our release through factories and automobiles every year of more than six billion tonnes of carbon dioxide, which helps air absorb heat from the sun, our atmosphere seems to be getting warmer.... [It's] been calculated that a few degrees rise in the earth's temperature would melt the polar ice caps."[41] Two years earlier, referencing work by Gilbert Plass, the *New York Times* had published an article arguing