Copyright

Hachette Books
Hachette Book Group
1290 Avenue of the Americas
New York, NY 10104
hachettebooks.com
twitter.com/hachettebooks

CONTENTS

Cover

Title Page

Copyright

Dedication

List of Figures

List of Tables

PART ONE: THE STAKES

Introduction

1. Standing at the Precipice
    - How We Got Here
    - Where We Might Go
    - The Precipice
2. Existential Risk
    - Understanding Existential Risk
    - Looking to the Present
    - Looking to Our Future
    - Looking to Our Past
    - Civilizational Virtues
    - Cosmic Significance
    - Uncertainty
    - Our Neglect of Existential Risks

PART TWO: THE RISKS

3. Natural Risks
    - Asteroids & Comets

Supervolcanic Eruptions
Stellar Explosions
Other Natural Risks
The Total Natural Risk

4. Anthropogenic Risks
Nuclear Weapons
Climate Change
Environmental Damage

5. Future Risks
Pandemics
Unaligned Artificial Intelligence
Dystopian Scenarios
Other Risks

PART THREE: THE PATH FORWARD

6. The Risk Landscape
Quantifying the Risks
Combining and Comparing Risks
Risk Factors
Which Risks?

7. Safeguarding Humanity
Grand Strategy for Humanity
Risks Without Precedent
International Coordination
Technological Progress
Research on Existential Risk
What You Can Do

8. Our Potential
Duration
Scale
Quality
Choices

Resources

Acknowledgments

Discover More

Appendices

Note on the Author

Note on the Type

Further Reading

Bibliography

Notes

To the hundred billion people before us, who fashioned our civilization;
To the seven billion now alive, whose actions may determine its fate;
To the trillions to come, whose existence lies in the balance.

LIST OF FIGURES

1.1 How we settled the world
1.2 The cradles of civilization
1.3 Striking improvements over the last 200 years
2.1 A classification of existential catastrophes
4.1 The number of stockpiled nuclear warheads over time
4.2 World population from 1700 to 2100
5.1 Measures of progress and interest in AI
5.2 An extended classification of existential catastrophes
6.1 How risks can combine
8.1 A timeline showing the scale of the past and future
D.1 How a 10% and 90% risk may combine

LIST OF TABLES

3.1 Progress in tracking near-Earth asteroids
3.2 The probability per century of a supervolcanic eruption
3.3 The probability per century of a stellar explosion
3.4 Estimates of total natural extinction risk via humanity's age
3.5 Estimates of total natural extinction risk via related species
3.6 The Big Five extinction events
4.1 Where is the carbon?
6.1 My existential risk estimates

PART ONE

THE STAKES

Introduction

If all goes well, human history is just beginning. Humanity is about two hundred thousand years old. But the Earth will remain habitable for hundreds of millions more—enough time for millions of future generations; enough to end disease, poverty and injustice forever; enough to create heights of flourishing unimaginable today. And if we could learn to reach out further into the cosmos, we could have more time yet: trillions of years, to explore billions of worlds. Such a lifespan places present-day humanity in its earliest infancy. A vast and extraordinary adulthood awaits.

Our view of this potential is easily obscured. The latest scandal draws our outrage; the latest tragedy, our sympathy. Time and space shrink. We forget the scale of the story in which we take part. But there are moments when we remember—when our vision shifts, and our priorities realign. We see a species precariously close to self-destruction, with a future of immense promise hanging in the balance. And which way that balance tips becomes our most urgent public concern.

This book argues that safeguarding humanity's future is the defining challenge of our time. For we stand at a crucial moment in the history of our species. Fueled by technological progress, our power has grown so great that for the first time in humanity's long history, we have the capacity to destroy ourselves—severing our entire future and everything we could become.

Yet humanity's wisdom has grown only falteringly, if at all, and lags dangerously behind. Humanity lacks the maturity, coordination and foresight necessary to avoid making mistakes from which we could never recover. As the gap between our power and our wisdom grows, our future is subject to an ever-increasing level of risk. This situation is unsustainable. So over the next few centuries, humanity will be tested: it will either act decisively to

protect itself and its longterm potential, or, in all likelihood, this will be lost forever.

To survive these challenges and secure our future, we must act now: managing the risks of today, averting those of tomorrow, and becoming the kind of society that will never pose such risks to itself again.

It is only in the last century that humanity's power to threaten its entire future became apparent. One of the most harrowing episodes has just recently come to light. On Saturday, October 27, 1962, a single officer on a Soviet submarine almost started a nuclear war. His name was Valentin Savitsky. He was captain of the submarine B-59—one of four submarines the Soviet Union had sent to support its military operations in Cuba. Each was armed with a secret weapon: a nuclear torpedo with explosive power comparable to the Hiroshima bomb.

It was the height of the Cuban Missile Crisis. Two weeks earlier, US aerial reconnaissance had produced photographic evidence that the Soviet Union was installing nuclear missiles in Cuba, from which they could strike directly at the mainland United States. In response, the US blockaded the seas around Cuba, drew up plans for an invasion and brought its nuclear forces to the unprecedented alert level of DEFCON 2 ("Next step to nuclear war").

On that Saturday, one of the blockading US warships detected Savitsky's submarine and attempted to force it to the surface by dropping low-explosive depth charges as warning shots. The submarine had been hiding deep underwater for days. It was out of radio contact, so the crew did not know whether war had already broken out. Conditions on board were extremely bad. It was built for the Arctic and its ventilator had broken in the tropical water. The heat inside was unbearable, ranging from 113°F near the torpedo tubes to 140°F in the engine room. Carbon dioxide had built up to dangerous concentrations, and crew members had begun to fall unconscious. Depth charges were exploding right next to the hull. One of the crew later recalled: "It felt like you were sitting in a metal barrel, which somebody is constantly blasting with a sledgehammer."

Increasingly desperate, Captain Savitsky ordered his crew to prepare their secret weapon:

> Maybe the war has already started up there, while we are doing somersaults here. We're going to blast them now! We will die, but we will sink them all—we will not disgrace our Navy![1]

Firing the nuclear weapon required the agreement of the submarine's political officer, who held the other half of the firing key. Despite the lack of authorization by Moscow, the political officer gave his consent.

On any of the other three submarines, this would have sufficed to launch their nuclear weapon. But by the purest luck, submarine B-59 carried the commander of the entire flotilla, Captain Vasili Arkhipov, and so required his additional consent. Arkhipov refused to grant it. Instead, he talked Captain Savitsky down from his rage and convinced him to give up: to surface amidst the US warships and await further orders from Moscow.[2]

We do not know precisely what would have happened if Arkhipov had granted his consent—or had he simply been stationed on any of the other three submarines. Perhaps Savitsky would not have followed through on his command. What is clear is that we came precariously close to a nuclear strike on the blockading fleet—a strike which would most likely have resulted in nuclear retaliation, then escalation to a full-scale nuclear war (the only kind the US had plans for). Years later, Robert McNamara, Secretary of Defense during the crisis, came to the same conclusion:

> No one should believe that had U.S. troops been attacked by nuclear warheads, the U.S. would have refrained from responding with nuclear warheads. Where would it have ended? In utter disaster.[3]

Ever since the advent of nuclear weapons, humans have been

making choices with such stakes. Ours is a world of flawed decision-makers, working with strikingly incomplete information, directing technologies which threaten the entire future of the species. We were lucky, that Saturday in 1962, and have so far avoided catastrophe. But our destructive capabilities continue to grow, and we cannot rely on luck forever.

We need to take decisive steps to end this period of escalating risk and safeguard our future. Fortunately, it is in our power to do so. The greatest risks are caused by human action, and they can be addressed by human action. Whether humanity survives this era is thus a choice humanity will make. But it is not an easy one. It all depends on how quickly we can come to understand and accept the fresh responsibilities that come with our unprecedented power.

This is a book about *existential risks*—risks that threaten the destruction of humanity's longterm potential. Extinction is the most obvious way humanity's entire potential could be destroyed, but there are others. If civilization across the globe were to suffer a truly unrecoverable collapse, that too would destroy our longterm potential. And we shall see that there are dystopian possibilities as well: ways we might get locked into a failed world with no way back.

While this set of risks is diverse, it is also exclusive. So I will have to set aside many important risks that fall short of this bar: our topic is not new dark ages for humanity or the natural world (terrible though they would be), but the permanent destruction of humanity's potential.

Existential risks present new kinds of challenges. They require us to coordinate globally and intergenerationally, in ways that go beyond what we have achieved so far. And they require foresight rather than trial and error. Since they allow no second chances, we need to build institutions to ensure that across our entire future we never once fall victim to such a catastrophe.

To do justice to this topic, we will have to cover a great deal of ground. Understanding the risks requires delving into physics, biology, earth science and computer science; situating this in the larger story of humanity requires history and anthropology;

discerning just how much is at stake requires moral philosophy and economics; and finding solutions requires international relations and political science. Doing this properly requires deep engagement with each of these disciplines, not just cherry-picking expert quotes or studies that support one's preconceptions. This would be an impossible task for any individual, so I am extremely grateful for the extensive advice and scrutiny of dozens of the world's leading researchers from across these fields.[4]

This book is ambitious in its aims. Through careful analysis of the potential of humanity and the risks we face, it makes the case that we live during the most important era of human history. Major risks to our entire future are a new problem, and our thinking has not caught up. So *The Precipice* presents a new ethical perspective: a major reorientation in the way we see the world, and our role in it. In doing so, the book aspires to start closing the gap between our wisdom and power, allowing humanity a clear view of what is at stake, so that we will make the choices necessary to safeguard our future.

I have not always been focused on protecting our longterm future, coming to the topic only reluctantly. I am a philosopher, at Oxford University, specializing in ethics. My earlier work was rooted in the more tangible concerns of global health and global poverty—in how we could best help the worst off. When coming to grips with these issues I felt the need to take my work in ethics beyond the ivory tower. I began advising the World Health Organization, World Bank and UK government on the ethics of global health. And finding that my own money could do hundreds of times as much good for those in poverty as it could do for me, I made a lifelong pledge to donate at least a tenth of all I earn to help them.[5] I founded a society, *Giving What We Can*, for those who wanted to join me, and was heartened to see thousands of people come together to pledge more than £1 billion over our lifetimes to the most effective charities we know of, working on the most important causes. Together, we've already been able to transform the lives of tens of thousands of people.[6] And because there are many other ways beyond our donations in which we can help fashion a better world, I helped start a wider movement, known as *effective altruism*, in which people aspire to use evidence and reason

to do as much good as possible.

Since there is so much work to be done to fix the needless suffering in our present, I was slow to turn to the future. It was so much less visceral; so much more abstract. Could it really be as urgent a problem as suffering now? As I reflected on the evidence and ideas that would culminate in this book, I came to realize that the risks to humanity's future are just as real and just as urgent—yet even more neglected. And that the people of the future may be even more powerless to protect themselves from the risks we impose than the dispossessed of our own time.

Addressing these risks has now become the central focus of my work: both researching the challenges we face, and advising groups such as the UK Prime Minister's Office, the World Economic Forum and DeepMind on how they can best address these challenges. Over time, I've seen a growing recognition of these risks, and of the need for concerted action.

To allow this book to reach a diverse readership, I've been ruthless in stripping out the jargon, needless technical detail and defensive qualifications typical of academic writing (my own included). Readers hungry for further technical detail or qualifications can delve into the many endnotes and appendices, written with them in mind.[7]

I have tried especially hard to examine the evidence and arguments carefully and even-handedly, making sure to present the key points even if they cut against my narrative. For it is of the utmost importance to get to the truth of these matters—humanity's attention is scarce and precious, and must not be wasted on flawed narratives or ideas[8].

Each chapter of *The Precipice* illuminates the central questions from a different angle. Part One (The Stakes) starts with a bird's-eye view of our unique moment in history, then examines why it warrants such urgent moral concern. Part Two (The Risks) delves into the science of the risks facing humanity, both from nature and from ourselves, showing that while some have been overstated, there is real risk and it is growing. So Part Three (The Path Forward) develops tools for understanding how these risks compare and combine, and new strategies for addressing them. I close with a vision of our future: of what we could achieve were we

to succeed.

This book is not just a familiar story of the perils of climate change or nuclear war. These risks that first awoke us to the possibilities of destroying ourselves are just the beginning. There are emerging risks, such as those arising from biotechnology and advanced artificial intelligence, that may pose much greater risk to humanity in the coming century.

Finally, this is not a pessimistic book. It does not present an inevitable arc of history culminating in our destruction. It is not a morality tale about our technological hubris and resulting fall. Far from it. The central claim is that there are real risks to our future, but that our choices can still make all the difference. I believe we are up to the task: that through our choices we can pull back from the precipice and, in time, create a future of astonishing value—with a richness of which we can barely dream, made possible by innovations we are yet to conceive. Indeed, my deep optimism about humanity's future is core to my motivation in writing this book. Our potential is vast. We have so much to protect.

1

Standing at the Precipice

It might be a familiar progression, transpiring on many worlds—a planet, newly formed, placidly revolves around its star; life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges which, at least up to a point, confers enormous survival value; and then technology is invented. It dawns on them that there are such things as laws of Nature, that these laws can be revealed by experiment, and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others, not so lucky or so prudent, perish.

—Carl Sagan[1]

We live at a time uniquely important to humanity's future. To see why, we need to take a step back and view the human story as a whole: how we got to this point and where we might be going next.

Our main focus will be humanity's ever-increasing power—power to improve our condition and power to inflict harm. We shall see how the major transitions in human history have enhanced our power, and enabled us to make extraordinary progress. If we can avoid catastrophe we can cautiously expect this progress to continue: the future of a responsible humanity is extraordinarily bright. But this increasing power has also brought on a new transition, at least as significant as any in our past, the transition to our time of perils.

Very little of humanity's story has been told; because very little *can* be told. Our species, *Homo sapiens*, arose on the savannas of Africa 200,000 years ago.[2] For an almost unimaginable time we have had great loves and friendships, suffered hardships and griefs, explored, created, and wondered about our place in the universe. Yet when we think of humanity's great achievements across time, we think almost exclusively of deeds recorded on clay, papyrus or paper—records that extend back only about 5,000 years. We rarely think of the first person to set foot in the strange new world of Australia some 70,000 years ago; of the first to name and study the plants and animals of each place we reached; of the stories, songs and poems of humanity in its youth.[3] But these accomplishments were real, and extraordinary.

We know that even before agriculture or civilization, humanity was a fresh force in the world. Using the simple, yet revolutionary, technologies of seafaring, clothing and fire, we traveled further than any mammal before us. We adapted to a wider range of environments, and spread across the globe.[4]

What made humanity exceptional, even at this nascent stage? We were not the biggest, the strongest or the hardiest. What set us apart was not physical, but mental—our intelligence, creativity and language.[5]

Yet even with these unique mental abilities, a single human alone in the wilderness would be nothing exceptional. He or she might be able to survive—intelligence making up for physical prowess—but would hardly dominate. In ecological terms, it is not a *human* that is remarkable, but *humanity*.

Each human's ability to cooperate with the dozens of other people in their band was unique among large animals. It allowed us to form something greater than ourselves. As our language grew in expressiveness and abstraction, we were able to make the most of such groupings: pooling together our knowledge, our ideas and our plans.

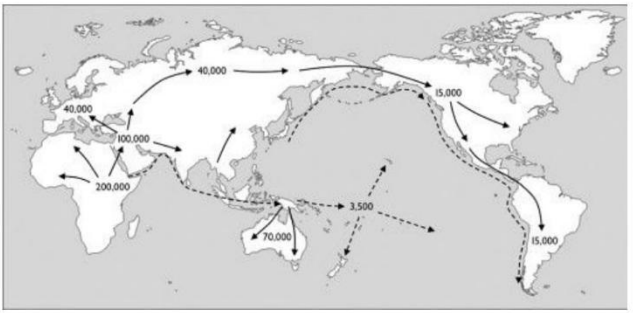

FIGURE 1.1 How we settled the world. The arrows show our current understanding of the land and sea routes taken by our ancestors, and how many years ago they reached each area.[6]

Crucially, we were able to cooperate across *time* as well as space. If each generation had to learn everything anew, then even a crude iron shovel would have been forever beyond our technological reach. But we learned from our ancestors, added minor innovations of our own, and passed this all down to our children. Instead of dozens of humans in cooperation, we had tens of thousands, cooperating across the generations, preserving and improving ideas through deep time. Little by little, our knowledge and our culture grew.[7]

At several points in the long history of humanity there has been a great transition: a change in human affairs that accelerated our accumulation of power and shaped everything that would follow. I will focus on three.[8]

The first was the Agricultural Revolution.[9] Around 10,000 years ago the people of the Fertile Crescent, in the Middle East, began planting wild wheat, barley, lentils and peas to supplement their foraging. By preferentially replanting the seeds from the best plants, they harnessed the power of evolution, creating new domesticated varieties with larger seeds and better yields. This worked with animals too, giving humans easier access to meat and hides, along with milk, wool and manure. And the physical power of draft animals to help plow the fields or transport the harvest was the biggest addition to humanity's power since fire.[10]

While the Fertile Crescent is often called "the cradle of civilization," in truth civilization had many cradles. Entirely independent agricultural revolutions occurred across the world in places where the climate and local species were suitable: in east Asia; sub-Saharan Africa; New Guinea; South, Central and North America; and perhaps elsewhere too.[11] The new practices fanned out from each of these cradles, changing the way of life for many from foraging to farming.

This had dramatic effects on the scale of human cooperation. Agriculture reduced the amount of land needed to support each person by a factor of a hundred, allowing large permanent settlements to develop, which began to unite together into states.[12] Where the largest foraging communities involved perhaps hundreds of people, some of the first cities had tens of thousands of inhabitants. At its height, the Sumerian civilization contained around a million people.[13] And 2,000 years ago, the Han dynasty of China reached sixty million people—about a *hundred thousand* times as many as were ever united in our forager past, and about ten times the entire global forager population at its peak.[14]

As more and more people were able to share their insights and discoveries, there were rapid developments in technology, institutions and culture. And the increasing numbers of people trading with one another made it possible for them to specialize in these areas—to devote a lifetime to governance, trade or the arts—allowing us to develop these ideas much more deeply.

Over the first 6,000 years of agriculture, we achieved world-changing breakthroughs including writing, mathematics, law and the wheel.[15] Of these, writing was especially important for strengthening our ability to cooperate across time and space: increasing the bandwidth between generations, the reliability of the information, and the distance over which ideas could be shared.

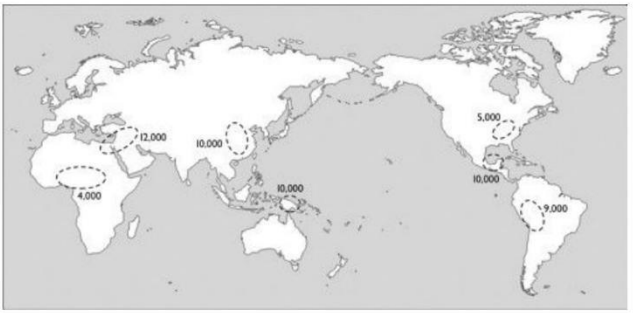

FIGURE 1.2 The cradles of civilization. The places around the world where agriculture was independently developed, marked with how many years ago this occurred.

The next great transition was the Scientific Revolution.[16] Early forms of science had been practiced since ancient times, and the seeds of empiricism can be found in the work of medieval scholars in the Islamic world and Europe.[17] But it was only about 400 years ago that humanity developed the scientific method and saw scientific progress take off.[18] This helped replace a reliance on received authorities with careful observation of the natural world, seeking simple and testable explanations for what we saw. The ability to test and discard bad explanations helped us break free from dogma, and allowed for the first time the systematic creation of knowledge about the workings of nature.

Some of our new-found knowledge could be harnessed to improve the world around us. So the accelerated accumulation of knowledge brought with it an acceleration of technological innovation, giving humanity increasing power over the natural world. The rapid pace allowed people to see transformative effects of these improvements within their own lifetimes. This gave rise to the modern idea of *progress*. Where the world had previously been dominated by narratives of decline and fall or of a recurring cycle, there was increasing interest in a new narrative: a grand project of working together to build a better future.

Soon, humanity underwent a third great transition: the Industrial Revolution. This was made possible by the discovery of immense reserves of energy in the form of coal and other fossil

fuels. These are formed from the compressed remains of organisms that lived in eons past, allowing us access to a portion of the sunlight that shone upon the Earth over millions of years.[19] We had already begun to drive simple machines with the renewable energy from the wind, rivers and forests; fossil fuels allowed access to vastly more energy, and in a much more concentrated and convenient form.

But energy is nothing without a way of converting it to useful work, to achieve our desired changes in the world. The steam engine allowed the stored chemical energy of coal to be turned into mechanical energy.[20] This mechanical energy was then used to drive machines that performed massive amounts of labor for us, allowing raw materials to be transformed into finished products much more quickly and cheaply than before. And via the railroad, this wealth could be distributed and traded across long distances.

Productivity and prosperity began to accelerate, and a rapid sequence of innovations ramped up the efficiency, scale and variety of automation, giving rise to the modern era of sustained economic growth.[21]

The effects of these transitions have not always been positive. Life in the centuries following the Agricultural Revolution generally involved more work, reduced nutrition and increased disease.[22] Science gave us weapons of destruction that haunt us to this day. And the Industrial Revolution was among the most destabilizing periods in human history. The unequal distribution of gains in prosperity and the exploitative labor practices led to the revolutionary upheavals of the early twentieth century.[23] Inequality between countries increased dramatically (a trend that has only begun to reverse in the last two decades).[24] Harnessing the energy stored in fossil fuels has released greenhouse gases, while industry fueled by this energy has endangered species, damaged ecosystems and polluted our environment.

Yet despite these real problems, on average human life today is substantially better than at any previous time. The most striking change may be in breaking free from poverty. Until 200 years ago —the last thousandth of our history[25]—increases in humanity's

power and prosperity came hand in hand with increases in the human population. Income *per person* stayed almost unchanged: a little above subsistence in times of plenty; a little below in times of need.[26] The Industrial Revolution broke this rule, allowing income to grow faster than population and ushering in an unprecedented rise in prosperity that continues to this day.

We often think of economic growth from the perspective of a society that is already affluent, where it is not immediately clear if further growth even improves our lives. But the most remarkable effects of economic growth have been for the poorest people. In today's world, one out of ten people are so poor that they live on less than two dollars per day—a widely used threshold for "extreme poverty." That so many have so little is among the greatest problems of our time, and has been a major focus of my life. It is shocking then to look further back and see that prior to the Industrial Revolution 19 out of 20 people lived on less than two dollars a day (even adjusting for inflation and purchasing power). Until the Industrial Revolution, any prosperity was confined to a tiny elite with extreme poverty the norm. But over the last two centuries more and more people have broken free from extreme poverty, and are now doing so more quickly than at any earlier time.[27] Two dollars a day is far from prosperity, and these statistics can be of little comfort to those who are still in the grip of poverty, but the trends toward improvement are clear.

And it is not only in terms of material conditions that life has improved. Consider education and health. Universal schooling has produced dramatic improvements in education. Before the Industrial Revolution, just one in ten of the world's people could read and write; now more than eight in ten can do so.[28] For the 10,000 years since the Agricultural Revolution, life expectancy had hovered between 20 and 30 years. It has now more than doubled, to 72 years.[29] And like literacy, these gains have been felt across the world. In 1800 the highest life expectancy of any country was a mere 43 years, in Iceland. Now every single country has a life expectancy above 50.[30] The industrial period has seen all of humanity become more prosperous, educated and long-lived than ever before. But we should not succumb to complacency in the face

of this astonishing progress. That we have achieved so much, and so quickly, should inspire us to address the suffering and injustices that remain.

We have also seen substantial improvements in our moral thinking.[31] One of the clearest trends is toward the gradual expansion of the moral community, with the recognition of the rights of women, children, the poor, foreigners and ethnic or religious minorities. We have also seen a marked shift away from violence as a morally acceptable part of society.[32] And in the last sixty years we have added the environment and the welfare of animals to our standard picture of morality. These social changes did not come naturally with prosperity. They were secured by reformers and activists, motivated by the belief that we can—and must—improve. We still have far to go before we are living up to these new ideals, and our progress can be painfully slow, but looking back even just one or two centuries shows how far we have come.

Of course, there have been many setbacks and exceptions. The path has been tumultuous, things have often become better in some ways while worse in others, and there is certainly a danger of choosing selectively from history to create a simple narrative of improvement from a barbarous past to a glorious present. Yet at the largest scales of human history, where we see not the rise and fall of each empire, but the changing face of human civilization across the entire globe, the trends toward progress are clear.[33]

It can be hard to believe such trends, when it so often feels like everything is collapsing around us. In part this skepticism comes from our everyday experience of our own lives or communities over a timespan of years—a scale where downs are almost as likely as ups. It might also come from our tendency to focus more on bad news than good and on threats rather than opportunities: heuristics that are useful for directing our actions, but which misfire when attempting to objectively assess the balance of bad and good.[34] When we try to overcome these distortions, looking for global indicators of the quality of our lives that are as objective as possible, it is very difficult to avoid seeing significant improvement from century to century.

And these trends should not surprise us. Every day we are the beneficiaries of uncountable innovations made by people over hundreds of thousands of years. Innovations in technology, mathematics, language, institutions, culture, art; the ideas of the hundred billion people who came before us, and shaped almost every facet of the modern world.[35] This is a stunning inheritance. No wonder, then, that our lives are better for it.

We cannot be sure these trends toward progress will continue. But given their tenacity, the burden would appear to be on the pessimist to explain why *now* is the point it will fail. This is especially true when people have been predicting such failure for so long and with such a poor track record. Thomas Macaulay made this point well:

> We cannot absolutely prove that those are in error who tell us that society has reached a turning point, that we have seen our best days. But so said all before us, and with just as much apparent reason... On what principle is it that, when we see nothing but improvement behind us, we are to expect nothing but deterioration before us?[36]

And he wrote those words in 1830, before an additional 190 years of progress and failed predictions of the end of progress. During those years, lifespan doubled, literacy soared and eight in ten people escaped extreme poverty. What might the coming years bring?

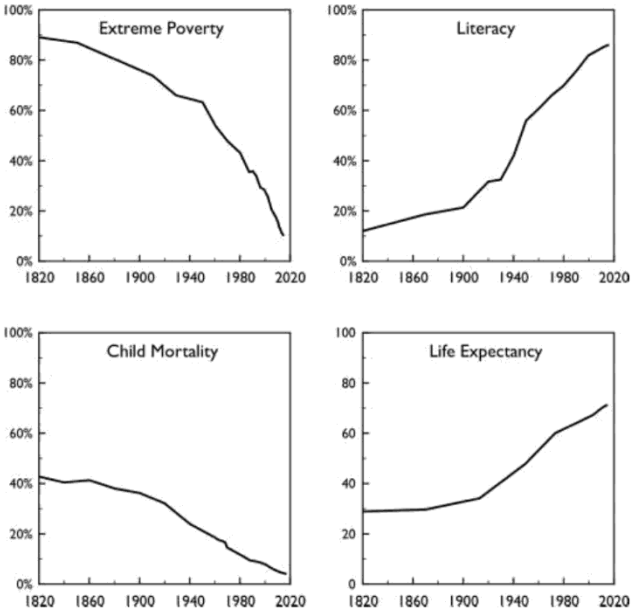

FIGURE 1.3 The striking improvements in extreme poverty, literacy, child mortality and life expectancy over the last 200 years.[37]

WHERE WE MIGHT GO

On the timescale of an individual human life, our 200,000-year history seems almost incomprehensibly long. But on a geological timescale it is short, and vanishingly so on the timescale of the universe as a whole. Our cosmos has a 14-billion-year history, and even that is short on the grandest scales. Trillions of years lie ahead of us. The future is immense.

How much of this future might we live to see? The fossil record provides some useful guidance. Mammalian species typically survive for around one million years before they go extinct; our close relative, *Homo erectus*, survived for almost two million.[38] If we think of one million years in terms of a single, eighty-year life, then today humanity would be in its adolescence—sixteen years old; just coming into our power; just old enough to get ourselves in serious trouble.[39]

Obviously, though, humanity is not a typical species. For one

thing, we have recently acquired a unique power to destroy ourselves—power that will be the focus of much of this book. But we also have unique power to protect ourselves from external destruction, and thus the potential to outlive our related species.

How long *could* we survive on Earth? Our planet will remain habitable for roughly a billion years.[40] That's enough time for trillions of human lives; time to watch mountain ranges rise, continents collide, orbits realign; and time, as well, to heal our society and our planet of the wounds we have caused in our immaturity.

And we might have more time yet. As one of the pioneers of rocketry put it, "Earth is the cradle of humanity, but one cannot live in a cradle forever."[41] We do not know, yet, how to reach other stars and settle their planets, but we know of no fundamental obstacles. The main impediment appears to be the time necessary to learn how. This makes me optimistic. After all, the first heavier-than-air flight was in 1903 and just sixty-eight years later we had launched a spacecraft that left our Solar System and will reach the stars. Our species learns quickly, especially in recent times, and a billion years is a long education. I think we will need far less.

If we can reach other stars, then the whole galaxy opens up to us. The Milky Way alone contains more than 100 billion stars, and some of these will last for trillions of years, greatly extending our potential lifespan. Then there are billions of other galaxies beyond our own. If we reach a future of such a scale, we might have a truly staggering number of descendants, with the time, resources, wisdom and experience to create a diversity of wonders unimaginable to us today.

While humanity has made progress toward greater prosperity, health, education and moral inclusiveness, there is so much further we could go. Our present world remains marred by malaria and HIV; depression and dementia; racism and sexism; torture and oppression. But with enough time, we can end these horrors—building a society that is truly just and humane.

And a world without agony and injustice is just a lower bound on how good life could be. Neither the sciences nor the humanities have yet found any upper bound. We get some hint at what is

possible during life's best moments: glimpses of raw joy, luminous beauty, soaring love. Moments when we are truly awake. These moments, however brief, point to possible heights of flourishing far beyond the status quo, and far beyond our current comprehension.

Our descendants could have eons to explore these heights, with new means of exploration. And it's not just wellbeing. Whatever you value—beauty, understanding, culture, consciousness, freedom, adventure, discovery, art—our descendants would be able to take these so much further, perhaps even discovering entirely new categories of value, completely unknown to us. Music we lack the ears to hear.

THE PRECIPICE

But this future is at risk. For we have recently undergone another transition in our power to transform the world—one at least as significant as the Agricultural, Scientific and Industrial Revolutions that preceded it.

With the detonation of the first atomic bomb, a new age of humanity began.[42] At that moment, our rapidly accelerating technological power finally reached the threshold where we might be able to destroy ourselves. The first point where the threat to humanity from within exceeded the threats from the natural world. A point where the entire future of humanity hangs in the balance. Where every advance our ancestors have made could be squandered, and every advance our descendants may achieve could be denied. The greater part of the book of human history left unwritten; the narrative broken off; blank pages.

Nuclear weapons were a discontinuous change in human power. At Hiroshima, a single bomb did the damage of thousands. And six years later, a single thermonuclear bomb held more energy than every explosive used in the entire course of the Second World War.[43]

It became clear that a war with such weapons would change the Earth in ways that were unprecedented in human history. World leaders, atomic scientists and public intellectuals began to

take seriously the possibility that a nuclear war would spell the end of humanity: either through extinction or a permanent collapse of civilization.[44] Early concern centered on radioactive fallout and damage to the ozone layer, but in the 1980s the focus shifted to a scenario known as nuclear winter, in which nuclear firestorms loft smoke from burning cities into the upper atmosphere.[45] High above the clouds, the smoke cannot be rained out and would persist for years, blackening the sky, chilling the Earth and causing massive crop failure. This was a mechanism by which nuclear war could result in extreme famine, not just in the combatant countries, but in every country around the world. Millions of direct deaths from the explosions could be followed by billions of deaths from starvation, and—potentially—by the end of humanity itself.

How close have we come to such a war? With so much to lose, nuclear war is in no one's interest. So we might expect these obvious dangers to create a certain kind of safety—where world leaders inevitably back down before the brink. But as more and more behind-the-scenes evidence from the Cold War has become public, it has become increasingly clear that we have only barely avoided full-scale nuclear war.

We saw how the intervention of a single person, Captain Vasili Arkhipov, may have prevented an all-out nuclear war at the height of the Cuban Missile Crisis. But even more shocking is just how many times in those few days we came close to disaster, only to be pulled back by the decisions of a few individuals.

The principal events of the crisis took place over a single week. On Monday, October 22, 1962, President John F. Kennedy gave a television address, informing his nation that the Soviets had begun installing strategic nuclear missiles in Cuba—directly threatening the United States. He warned that any use of these nuclear weapons would be met by a full-scale nuclear retaliation on the Soviet Union. His advisers drew up plans for both air strikes on the 48 missiles they had discovered and a full invasion of Cuba. US forces were brought to DEFCON 3, to prepare for a possible nuclear war.[46]

by the Soviet Union on the United States, requiring a full retaliatory response upon the Soviet Union."[50]

It is extremely difficult to estimate the chance that the crisis would have escalated to nuclear war.[51] Shortly after, Kennedy told a close adviser that he thought the probability of it ending in nuclear war with the USSR was "somewhere between one out of three, and even."[52] And it has just been revealed that the day after the crisis ended, Paul Nitze (an adviser to Kennedy's war council) estimated the chance at 10 percent, and thought that everyone else in the council would have put it even higher.[53] Moreover, none of these people knew about the tactical nuclear weapons in Cuba, Khrushchev's lack of control of his troops or the events on submarine B-59.

While I'm reluctant to question those whose very decisions could have started the war, my own view is that they were somewhat too pessimistic, given what they knew at the time. However, when we include the subsequent revelations about what was really happening in Cuba my estimates would roughly match theirs. I'd put the chance of the crisis escalating to a nuclear war with the Soviets at something between 10 and 50 percent.[54]

When writing about such close calls, there is a tendency to equate this chance to that of the end of civilization or the end of humanity itself. But that would be a large and needless exaggeration. For we need to combine this chance of nuclear war with the chance that such a war would spell the end of humanity or human civilization, which is far from certain. Yet even making such allowances the Cuban Missile Crisis would remain one of the pivotal moments in 200,000 years of human history: perhaps the closest we have ever come to losing it all.

Even now, with the Cold War just a memory, nuclear weapons still pose a threat to humanity. At the time of writing, the highest chance of a nuclear conflict probably involves North Korea. But not all nuclear wars are equal. North Korea has less than 1 percent as many warheads as Russia or the US, and they are substantially smaller. A nuclear war with North Korea would be a terrible disaster, but it currently poses little threat to humanity's longterm potential.[55]

Instead, most of the existential risk from nuclear weapons today probably still comes from the enormous American and Russian arsenals. The development of ICBMs (intercontinental ballistic missiles) allowed each side to destroy most of the other's missiles with just thirty minutes' warning, so they each moved many missiles to "hair-trigger alert"—ready to launch in just ten minutes.[56] Such hair-trigger missiles are extremely vulnerable to accidental launch, or to deliberate launch during a false alarm. As we shall see in Chapter 4, there has been a chilling catalog of false alarms continuing past the end of the Cold War. On a longer timescale there is also the risk of other nations creating their own enormous stockpiles, of innovations in military technologies undermining the logic of deterrence, and of shifts in the geopolitical landscape igniting another arms race between great powers.

Nuclear weapons are not the only threat to humanity. They have been our focus so far because they were the first major risk and have already threatened humanity. But there are others too.

The exponential rise in prosperity brought on by the Industrial Revolution came on the back of a rapid rise in carbon emissions. A minor side effect of industrialization has eventually grown to become a global threat to health, the environment, international stability, and maybe even humanity itself.

Nuclear weapons and climate change have striking similarities and contrasts. They both threaten humanity through major shifts in the Earth's temperature, but in opposite directions. One burst in upon the scene as the product of an unpredictable scientific breakthrough; the other is the continuation of centuries-long scaling-up of old technologies. One poses a small risk of sudden and precipitous catastrophe; the other is a gradual, continuous process, with a delayed onset—where some level of catastrophe is assured and the major uncertainty lies in just how bad it will be. One involves a classified military technology controlled by a handful of powerful actors; the other involves the aggregation of small effects from the choices of everyone in the world.

As technology continues to advance, new threats appear on

the horizon. These threats promise to be more like nuclear weapons than like climate change: resulting from sudden breakthroughs, precipitous catastrophes, and the actions of a small number of actors. There are two emerging technologies that especially concern me; they will be the focus of Chapter 5.

Ever since the Agricultural Revolution, we have induced genetic changes in the plants and animals around us to suit our ends. But the discovery of the genetic code and the creation of tools to read and write it have led to an explosion in our ability to refashion life to new purposes. Biotechnology will bring major improvements in medicine, agriculture and industry. But it will also bring risks to civilization and to humanity itself: both from accidents during legitimate research and from engineered bioweapons.

We are also seeing rapid progress in the capabilities of artificial intelligence (AI) systems, with the biggest improvements in the areas where AI has traditionally been weakest, such as perception, learning and general intelligence. Experts find it likely that this will be the century that AI exceeds human ability not just in a narrow domain, but in general intelligence—the ability to overcome a diverse range of obstacles to achieve one's goals. Humanity has risen to a position where we control the rest of the world precisely because of our unparalleled mental abilities. If we pass this mantle to our machines, it will be they who are in this unique position. This should give us cause to wonder why it would be humanity who will continue to call the shots. We need to learn how to align the goals of increasingly intelligent and autonomous machines with human interests, and we need to do so before those machines become more powerful than we are.

These threats to humanity, and how we address them, define our time. The advent of nuclear weapons posed a real risk of human extinction in the twentieth century. With the continued acceleration of technology, and without serious efforts to protect humanity, there is strong reason to believe the risk will be higher this century, and increasing with each century that technological progress continues. Because these anthropogenic risks outstrip all natural risks combined, they set the clock on how long humanity

has left to pull back from the brink.

I am not claiming that extinction is the inevitable conclusion of scientific progress, or even the most likely outcome. What I am claiming is that there has been a robust trend toward increases in the power of humanity which has reached a point where we pose a serious risk to our own existence. How we react to this risk is up to us.

Nor am I arguing against technology. Technology has proved itself immensely valuable in improving the human condition. And technology is essential for humanity to achieve its longterm potential. Without it, we would be doomed by the accumulated risk of natural disasters such as asteroid impacts. Without it, we would never achieve the highest flourishing of which we are capable.

The problem is not so much an excess of technology as a lack of wisdom.[57] Carl Sagan put this especially well:

> Many of the dangers we face indeed arise from science and technology—but, more fundamentally, because we have become powerful without becoming commensurately wise. The world-altering powers that technology has delivered into our hands now require a degree of consideration and foresight that has never before been asked of us.[58]

This idea has even been advocated by a sitting US president:

> the very spark that marks us as a species—our thoughts, our imagination, our language, our tool-making, our ability to set ourselves apart from nature and bend it to our will—those very things also give us the capacity for unmatched destruction... Technological progress without an equivalent progress in human institutions can doom us. The scientific revolution that led to the splitting of an atom requires a moral revolution as well.[59]

We need to gain this wisdom; to have this moral revolution. Because we cannot come back from extinction, we cannot wait

until a threat strikes before acting—we must be proactive. And because gaining wisdom or starting a moral revolution takes time, we need to start now.

I think that we are likely to make it through this period. Not because the challenges are small, but because we will rise to them. The very fact that these risks stem from human action shows us that human action can address them.[60] Defeatism would be both unwarranted and counterproductive—a self-fulfilling prophecy. Instead, we must address these challenges head-on with clear and rigorous thinking, guided by a positive vision of the longterm future we are trying to protect.

How big are these risks? One cannot expect precise numbers, as the risks are *complex* (so not amenable to simple mathematical analysis) and *unprecedented* (so cannot be approximated by a longterm frequency). Yet it is important to at least try to give quantitative estimates. Qualitative statements such as "a grave risk of human extinction" could be interpreted as meaning anything from 1 percent all the way to 99 percent.[61] They add more confusion than clarity. So I will offer quantitative estimates, with the proviso that they can't be precise and are open to revision.

During the twentieth century, my best guess is that we faced around a one in a hundred risk of human extinction or the unrecoverable collapse of civilization. Given everything I know, I put the existential risk this century at around one in six: Russian roulette.[62] (See table 6.1 here for a breakdown of the risks.) If we do not get our act together, if we continue to let our growth in power outstrip that of wisdom, we should expect this risk to be even higher next century, and each successive century.

These are the greatest risks we have faced.[63] If I'm even roughly right about their scale, then we cannot survive many centuries with risk like this. It is an *unsustainable* level of risk.[64] Thus, one way or another, this period is unlikely to last more than a small number of centuries.[65] Either humanity takes control of its destiny and reduces the risk to a sustainable level, or we destroy ourselves.

Consider human history as a grand journey through the wilderness. There are wrong turns and times of hardship, but also

2

Existential Risk

> *The crucial role we fill, as moral beings, is as members of a cross-generational community, a community of beings who look before and after, who interpret the past in light of the present, who see the future as growing out of the past, who see themselves as members of enduring families, nations, cultures, traditions.*
>
> —Annette Baier[1]

We have seen how the long arc of human history has brought us to a uniquely important time in our story—a period where our entire future hangs in the balance. And we have seen a little of what might lie beyond, if only we can overcome the risks.

It is now time to think more deeply about what is at stake; why safeguarding humanity through this time is so important. To do so we first need to clarify the idea of existential risk. What exactly *is* existential risk? How does it relate to more familiar ideas of extinction or the collapse of civilization?

We can then ask what it is about these risks that compels our urgent concern. The chief reason, in my view, is that we would lose our entire future: everything humanity could be and everything we could achieve. But that is not all. The case that it is crucial to safeguard our future draws support from a wide range of moral traditions and foundations. Existential risks also threaten to destroy our present, and to betray our past. They test civilization's virtues, and threaten to remove what may be the most complex and significant part of the universe.

If we take any of these reasons seriously, we have a lot of work to do to protect our future. For existential risk is greatly neglected:

by government, by academia, by civil society. We will see why this has been the case, and why there is good reason to suspect this will change.

UNDERSTANDING EXISTENTIAL RISK

Humanity's future is ripe with possibility. We have achieved a rich understanding of the world we inhabit and a level of health and prosperity of which our ancestors could only dream. We have begun to explore the other worlds in the heavens above us, and to create virtual worlds completely beyond our ancestors' comprehension. We know of almost no limits to what we might ultimately achieve.

Human extinction would foreclose our future. It would destroy our potential. It would eliminate all possibilities but one: a world bereft of human flourishing. Extinction would bring about this failed world and lock it in forever—there would be no coming back.

The philosopher Nick Bostrom showed that extinction is not the only way this could happen: there are other catastrophic outcomes in which we lose not just the present, but all our potential for the future.[2]

Consider a world in ruins: an immense catastrophe has triggered a global collapse of civilization, reducing humanity to a pre-agricultural state. During this catastrophe, the Earth's environment was damaged so severely that it has become impossible for the survivors to ever re-establish civilization. Even if such a catastrophe did not cause our extinction, it would have a similar effect on our future. The vast realm of futures currently open to us would have collapsed to a narrow range of meager options. We would have a failed world with no way back.

Or consider a world in chains: in a future reminiscent of George Orwell's *Nineteen Eighty-Four*, the entire world has become locked under the rule of an oppressive totalitarian regime, determined to perpetuate itself. Through powerful, technologically enabled indoctrination, surveillance and enforcement, it has become impossible for even a handful of dissidents to find each other, let alone stage an uprising. With everyone on Earth living under such rule, the regime is stable from

threats, internal and external. If such a regime could be maintained indefinitely, then descent into this totalitarian future would also have much in common with extinction: just a narrow range of terrible futures remaining, and no way out.

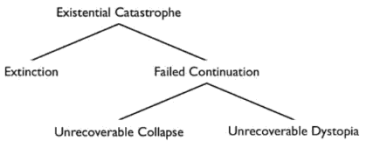

FIGURE 2.1 A classification of existential catastrophes by the kind of outcome that gets locked in.

Following Bostrom, I shall call these "existential catastrophes," defining them as follows:[3]

> An *existential catastrophe* is the destruction of humanity's longterm potential.
>
> An *existential risk* is a risk that threatens the destruction of humanity's longterm potential.

These definitions capture the idea that the outcome of an existential catastrophe is both dismal and irrevocable. We will not just fail to fulfill our potential, but this very potential itself will be permanently lost. While I want to keep the official definitions succinct, there are several areas that warrant clarification.

First, I am understanding *humanity's longterm potential* in terms of the set of all possible futures that remain open to us.[4] This is an expansive idea of possibility, including everything that humanity could eventually achieve, even if we have yet to invent the means of achieving it.[5] But it follows that while our choices can lock things in, closing off possibilities, they can't open up new ones. So any reduction in humanity's potential should be understood as permanent. The challenge of our time is to *preserve* our vast potential, and to *protect* it against the risk of future destruction. The ultimate purpose is to allow our descendants to *fulfill* our potential, realizing one of the best possible futures open to us.

While it may seem abstract at this scale, this is really a familiar

idea that we encounter every day. Consider a child with high longterm potential: with futures open to her in which she leads a great life. It is important that her potential is preserved: that her best futures aren't cut off due to accident, trauma or lack of education. It is important that her potential is protected: that we build in safeguards to make such a loss of potential extremely unlikely. And it is important that she ultimately fulfills her potential: that she ends up taking one of the best paths open to her. So too for humanity.

Existential risks threaten the destruction of humanity's potential. This includes cases where this destruction is complete (such as extinction) and where it is nearly complete, such as a permanent collapse of civilization in which the possibility for some very minor types of flourishing remain, or where there remains some remote chance of recovery.[6] I leave the thresholds vague, but it should be understood that in any existential catastrophe the greater part of our potential is gone and very little remains.[7]

Second, my focus on humanity in the definitions is not supposed to exclude considerations of the value of the environment, other animals, successors to *Homo sapiens*, or creatures elsewhere in the cosmos. It is not that I think only humans count. Instead, it is that humans are the only beings we know of that are responsive to moral reasons and moral argument —the beings who can examine the world and decide to do what is best. If we fail, that upward force, that capacity to push toward what is best or what is just, will vanish from the world.

Our potential is a matter of what humanity can achieve through the combined actions of each and every human. The value of our actions will stem in part from what we do to and for humans, but it will depend on the effects of our actions on non-humans too. If we somehow give rise to new kinds of moral agents in the future, the term "humanity" in my definition should be taken to include them.

My focus on humanity prevents threats to a single country or culture from counting as existential risks. There is a similar term that gets used this way—when people say that something is "an existential threat to this country." Setting aside the fact that these claims are usually hyperbole, they are expressing a similar idea:

that something threatens to permanently destroy the longterm potential of a country or culture.[8] However, it is very important to keep talk of an "existential risk" (without any explicit restriction to a group) to apply only to threats against the whole of humanity.

Third, any notion of risk must involve some kind of probability. What kind is involved in existential risk? Understanding the probability in terms of objective long-run frequencies won't work, as the existential catastrophes we are concerned with can only ever happen once, and will always be unprecedented until the moment it is too late. We can't say the probability of an existential catastrophe is precisely zero just because it hasn't happened *yet*.

Situations like these require an evidential sense of probability, which describes the appropriate degree of belief we should have on the basis of the available information. This is the familiar type of probability used in courtrooms, banks and betting shops. When I speak of the probability of an existential catastrophe, I will mean the credence humanity should have that it will occur, in light of our best evidence.[9]

There are many utterly terrible outcomes that do not count as existential catastrophes.

One way this could happen is if there were no single precipitous event, but a multitude of smaller failures. This is because I take on the usual sense of catastrophe as a single, decisive event, rather than any combination of events that is bad in sum. If we were to squander our future simply by continually treating each other badly, or by never getting around to doing anything great, this could be just as bad an outcome but wouldn't have come about via a catastrophe.

Alternatively, there might be a single catastrophe, but one that leaves open some way for humanity to eventually recover. From our own vantage, looking out to the next few generations, this may appear equally bleak. But a thousand years hence it may be considered just one of several dark episodes in the human story. A true existential catastrophe must by its very nature be the decisive moment of human history—the point where we failed.

build a deflection system, or to ignore the issue and run the risk. To the contrary, responding to the threat would immediately become one of the world's top priorities. Thus our lack of concern about these threats is much more to do with not yet believing that there are such threats, than it is about seriously doubting the immensity of the stakes.

Yet it is important to spend a little while trying to understand more clearly the different sources of this importance. Such an understanding can buttress feeling and inspire action; it can bring to light new considerations; and it can aid in decisions about how to set our priorities.

LOOKING TO THE PRESENT

Not all existential catastrophes involve human extinction, and not all methods of extinction involve pain or untimely death. For example, it is theoretically possible that we could all simply decide not to reproduce. This could destroy our potential without, let us suppose, causing any suffering. But the existential risks we actually face are not so peaceful. Rather, they are obviously horrific by the most familiar moral standards.

If, over the coming century, humanity is destroyed in a nuclear winter, or an engineered pandemic, or a catastrophic war involving some new technology, then seven billion lives would be cut short—including, perhaps, your own life, or the lives of those you love. Many would likely die in agony—starving, or burning, or ravaged by disease.

The moral case for preventing such horror needs little elaboration. Humanity has seen catastrophes before, on smaller scales: thousands, or millions, of human lives destroyed. We know how tremendously important it is to prevent such disasters. At such a scale, we lose our ability to fully comprehend the magnitude of what is lost, but even then the numbers provide a guide to the moral stakes.[18] Other things being equal, millions of deaths must be much worse than thousands of deaths; and billions, much worse than millions. Even measured just in terms of lives cut short, human extinction would easily be the worst event in our long history.

But an existential catastrophe is not just a catastrophe that destroys a particularly large number of lives. It destroys our potential.

My mentor, Derek Parfit, asked us to imagine a devastating nuclear war killing 99 percent of the world's people.[19] A war that would leave behind a dark age lasting centuries, before the survivors could eventually rebuild civilization to its former heights; humbled, scarred—but undefeated.

Now compare this with a war killing a full 100 percent of the world's people. This second war would be worse, of course, but how much worse? Either war would be the worst catastrophe in history. Either would kill billions. The second war would involve tens of millions of additional deaths, and so would be worse for this reason. But there is another, far more significant difference between the two wars. Both wars kill billions of humans; but the second war kills humanity. Both wars destroy our present; but the second war destroys our future.

It is this qualitative difference in what is lost with that last percent that makes existential catastrophes unique, and that makes reducing the risk of existential catastrophe uniquely important.[20]

In expectation, almost all humans who will ever live have yet to be born. Absent catastrophe, most generations are future generations. As the writer Jonathan Schell put it:

> The procession of generations that extends onwards from our present leads far, far beyond the line of our sight, and, compared with these stretches of human time, which exceed the whole history of the earth up to now, our brief civilized moment is almost infinitesimal. Yet we threaten, in the name of our transient aims and fallible convictions, to foreclose it all. If our species does destroy itself, it will be a death in the cradle—a case of infant mortality.[21]

And because, in expectation, almost all of humanity's life lies in

the future, almost everything of value lies in the future as well: almost all the flourishing; almost all the beauty; our greatest achievements; our most just societies; our most profound discoveries.[22] We can continue our progress on prosperity, health, justice, freedom and moral thought. We can create a world of wellbeing and flourishing that challenges our capacity to imagine. And if we protect that world from catastrophe, it could last millions of centuries. This is our potential—what we could achieve if we pass the Precipice and continue striving for a better world.

It is this view of the future—the immense value of humanity's potential—that most persuades me to focus my energies on reducing existential risk. When I think of the millions of future generations yet to come, the importance of protecting humanity's future is clear to me. To risk destroying this future, for the sake of some advantage limited only to the present, seems to me profoundly parochial and dangerously short-sighted. Such neglect privileges a tiny sliver of our story over the grand sweep of the whole; it privileges a tiny minority of humans over the overwhelming majority yet to be born; it privileges this particular century over the millions, or maybe billions, yet to come.[23]

To see why this would be wrong, consider an analogy with distance. A person does not matter less, the further away from you they are in space. It matters just as much if my wife gets sick while she is away at a conference in Kenya as if she gets sick while home with me in Oxford. And the welfare of strangers in Kenya matters just as much as the welfare of strangers in Oxford. Of course, we may have special duties to some individuals—to family; to members of the same community—but it is never spatial distance, in itself, that determines these differences in our obligations. Recognizing that people matter equally, regardless of their geographic location, is a crucial form of moral progress, and one that we could do much more to integrate into our policies and our philanthropy.

People matter equally regardless of their temporal location too. Our lives matter just as much as those lived thousands of years ago, or those a thousand years hence.[24] Just as it would be wrong to think that other people matter less the further they are from you in space, so it is to think they matter less the further away

from you they are in time. The value of their happiness, and the horror of their suffering, is undiminished.

Recognizing that people matter equally, wherever they are in time, is a crucial next step in the ongoing story of humanity's moral progress. Many of us recognize this equality to some extent already. We know it is wrong to make future generations worse off in order to secure lesser benefits for ourselves. And if asked, we would agree that people now don't objectively matter more than people in the future. But we assume that this leaves most of our priorities unaltered. For example, thinking that long-run effects of our choices quickly disappear; that they are so uncertain that the good cancels the bad; or that people in the future will be much better situated to help themselves.[25]

But the possibility of preventable existential risks in our lifetimes shows that there are issues where our actions can have sustained positive effects over the whole longterm future, and where we are the only generation in a position to produce those effects.[26] So the view that people in the future matter just as much as us has deep practical implications. We have a long way to go if we are to understand these and integrate them fully into our moral thinking.

Considerations like these suggest an ethic we might call *longtermism*, which is especially concerned with the impacts of our actions upon the longterm future.[27] It takes seriously the fact that our own generation is but one page in a much longer story, and that our most important role may be how we shape—or fail to shape—that story. Working to safeguard humanity's potential is one avenue for such a lasting impact and there may be others too.[28]

One doesn't have to approach existential risk from this direction—there is already a strong moral case just from the immediate effects—but a longtermist ethic is nevertheless especially well suited to grappling with existential risk. For longtermism is animated by a moral re-orientation toward the vast future that existential risks threaten to foreclose.

Of course, there are complexities.

When economists evaluate future benefits, they use a method called discounting, which dampens ("discounts") benefits based on how far away they are in time. If one took a commonly used discount rate of 5 percent per year and applied it to our future, there would be strikingly little value left. Applied naïvely, this discount rate would suggest our entire future is worth only about twenty times as much as the coming year, and that the period from 2100 to eternity is worth less than the coming year. Does this call into question the idea that our future is extremely valuable?

No. Results like this arise only from an incorrect application of the economic methods. When the subtleties of the problem are taken into account and discounting is correctly applied, the future is accorded an extremely high value. The mathematical details would take us too far afield, but for now it suffices to note that discounting human wellbeing (as opposed to instrumental goods such as money), purely on the basis of distance away from us in time, is deeply implausible—especially over the long time periods we are discussing. It implies, for example, that if you can save one person from a headache in a million years' time, or a billion people from torture in two million years, you should save the one from a headache.[29] A full explanation of why economic discounting does not trivialize the value of the longterm future can be found in Appendix A.

Some philosophers question the value of protecting our longterm future for quite a different reason. They note that the timing of the benefits is not the only unusual feature of this case. If we save humanity from extinction, that will change the number of people who will ever live. This brings up ethical issues that don't arise when simply saving the lives of existing people. Some of the more extreme approaches to this relatively new field of "population ethics" imply that there is no reason to avoid extinction stemming from considerations of future generations—it just doesn't matter whether these future people come into being or not.

A full treatment of these matters would take too long and be of interest only to a few, so I reserve the detailed discussion for Appendix B. To briefly summarize: I do not find these views very plausible, either. They struggle to capture our reasons to care

> many generations, it becomes a partnership not only between those who are living, but between those who are living, those who are dead, and those who are to be born.[35]

This might give us reasons to safeguard humanity that are grounded in our past—obligations to our grandparents, as well as our grandchildren.

Our ancestors set in motion great projects for humanity that are too big for any single generation to achieve. Projects such as bringing an end to war, forging a just world and understanding our universe. In the year 65 CE, Seneca the Younger explicitly set out such a vast intergenerational project:

> The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Let us be satisfied with what we have found out, and let our descendants also contribute something to the truth... Many discoveries are reserved for ages still to come, when memory of us will have been effaced.[36]

It is astounding to be spoken to so directly across such a gulf of time, and to see this 2,000-year plan continue to unfold.[37]

A human, or an entire generation, cannot complete such grand projects. But humanity can. We work together, each generation making a little progress while building up the capacities, resources and institutions to empower future generations to take the next step.

Indeed, when I think of the unbroken chain of generations leading to our time and of everything they have built for us, I am humbled. I am overwhelmed with gratitude; shocked by the enormity of the inheritance and at the impossibility of returning

even the smallest fraction of the favor. Because a hundred billion of the people to whom I owe everything are gone forever, and because what they created is so much larger than my life, than my entire generation.

The same is true at the personal level. In the months after my daughter was born, the magnitude of everything my parents did for me was fully revealed. I was shocked. I told them; thanked them; apologized for the impossibility of ever repaying them. And they smiled, telling me that this wasn't how it worked—that one doesn't repay one's parents. One passes it on.

My parents aren't philosophers. But their remarks suggest another way in which the past could ground our duties to the future. Because the arrow of time makes it so much easier to help people who come after you than people who come before, the best way of understanding the partnership of the generations may be asymmetrical, with duties all flowing forwards in time—paying it forwards. On this view, our duties to future generations may thus be grounded in the work our ancestors did for us when *we* were future generations.[38]

So if we drop the baton, succumbing to an existential catastrophe, we would fail our ancestors in a multitude of ways. We would fail to achieve the dreams they hoped for; we would betray the trust they placed in us, their heirs; and we would fail in any duty we had to pay forward the work they did for us. To neglect existential risk might thus be to wrong not only the people of the future, but the people of the past.

It would also be to risk the destruction of everything of value from the past we might have reason to preserve.[39] Some philosophers have suggested that the right way to respond to some valuable things is not to promote them, but to protect or preserve them; to cherish or revere them.[40] We often treat the value of cultural traditions in this way. We see indigenous languages and ways of life under threat—perhaps to be lost forever to this world—and we are filled with a desire to preserve them, and protect them from future threats.

Someone who saw the value of humanity in this light may not be so moved by the loss of what could have been. But they would still be horrified by extinction: the ruin of every cathedral and