Morality, Foresight, and Human Flourishing

"Thinking about existential risks is neither fun, nor easy. It is also fraught with risk itself. When it comes to technology developments, the ones with the greatest impact are usually the ones that are the most unanticipated. Nevertheless, as Louis Pasteur said, 'fortune favors the prepared mind,' and unless we try and prepare as carefully as we can for a future in which technology evolves at an exponential rate, the likelihood that the future could bring catastrophe on a global scale will increase. This book presents a sober and careful examination of the emerging field of existential risk studies, and will provide a useful introduction to all those who want to come up to speed quickly on developments over the past decade."

—**Lawrence M. Krauss**, Director of the Origins Project at Arizona State University, and Chair of the Board of Sponsors of the *Bulletin of the Atomic Scientists*. His most recent book is *The Greatest Story Ever Told… So Far: Why are we here?*

"*Morality, Foresight, and Human Flourishing* is an excellent introduction to a new and important area of research. I hope it will be widely read."

—**Peter Singer**, Ira W. DeCamp Professor of Bioethics at Princeton University and author of *Animal Liberation* and *The Most Good You Can Do*

"The path to our future is rife with threats to the very existence of humanity. How can we avoid creating technologies that destroy us, as well as other global catastrophes? We need a roadmap, and this is precisely what Torres provides in this carefully thought-out and useful book."

—**Susan Schneider**, Associate Professor at the University of

Pitchstone Publishing
Durham, North Carolina
www.pitchstonepublishing.com

First edition

Printed in the USA

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Names: Torres, Phil, author.
Title: Morality, foresight, and human flourishing : an introduction to existential risks / Phil Torres ; foreword by Martin Rees.
Description: Durham, North Carolina : Pitchstone Publishing, [2017] | Includes bibliographical references and index.
Identifiers: LCCN 2017024484| ISBN 9781634311427 (pbk. : alk. paper) | ISBN 9781634311441 (epdf) | ISBN 9781634311458 (mobi)
Subjects: LCSH: Risk—Sociological aspects. | Natural disasters. | Environmental disasters. | Technology—Risk assessment.
Classification: LCC HM1101 .T67 2017 | DDC 363.34—dc23
LC record available at https://lccn.loc.gov/2017024484

Contents

Foreword

This is a welcome and timely book that draws attention to issues that our civilization's entire fate may depend on—and that need far more study and focus than they currently receive.

Our Earth is 45 million centuries old. But this century is the first when one species—ours—can determine the biosphere's fate. We're deep in a new era called the Anthropocene, where the main threats come not from nature, but from ourselves. In the crises of the Cold War era, the probability of stumbling toward Armageddon was put by some as high as one in three. That's tens of thousands of times higher than for an equally catastrophic asteroid impact.

Those of us with cushioned lives in the developed world fret too much about improbable air crashes, carcinogens in food, low radiation doses, and so forth. Current terrorism disproportionately fills the headlines. But we're in denial about far more shattering scenarios that thankfully haven't yet happened, but could.

The "x-risks" that threaten us are of two kinds. First, a growing population, more demanding of food, energy, and other natural resources, is putting unsustainable pressure on ecosystems, threatening loss of biodiversity and the crossing of climatic "tipping points."

But there's a second class of threats that will loom even larger: those stemming from the misuse, by error or design, of ever more powerful technologies. Nuclear weapons are based on twentieth-century science. But twenty-first-century sciences—biotech, cybertech, and artificial intelligence (AI)—will pose risks that are even more intractable.

Advances in genetics and microbiology offer exciting prospects,

but they have downsides. It's accepted that techniques like "gain-of-function" modification of viruses and CRISPR/Cas9 gene editing will need regulation. There are precedents here: in the early days of recombinant DNA research, a group of biologists formulated the Asilomar Declaration, setting up guidelines on what experiments should and shouldn't be done. In the same spirit there's a call for similar regulation of the new techniques. However, the research community today, 40 years after Asilomar, is far larger, far more broadly international, and far more influenced by commercial pressures. Whatever regulations are imposed, on prudential or ethical grounds, could never be fully enforced worldwide—any more than the drug laws or tax laws can. Whatever can be done will be done by someone, somewhere. And that is deeply scary.

In consequence, maybe the most intractable challenges to all governments will stem from the rising empowerment of tech-savvy groups (or even individuals), by bio-as well as cybertechnology. This will aggravate the tension between freedom, privacy, and security.

These bio-concerns are relatively near-term—within 10 or 15 years. What about robotics and AI? Cyber threats are of course already pervasive and costly. And though we don't yet have the human-level robots that have been a staple of science fiction for decades, some experts think they will one day be real. If they could infiltrate the Internet—and the Internet of things—they could manipulate the rest of the world. They may have goals utterly orthogonal to human wishes—or even treat humans as an encumbrance. So how can we ensure that ever more sophisticated computers remain docile "idiot savants" and don't "go rogue"?

Experts disagree on how long it will take before machines achieve general-purpose human-level intelligence. Some say 25 years. Others say never. The median guess in a recent survey was about 50 years. And it's claimed that once a threshold is crossed, there will be an intelligence explosion. That's because electronics is a million times faster than the transmission of signals in the brain, and because

computers can network and exchange information much faster than we can by speaking.

There is perhaps a parallel with nuclear fusion. Making an explosion—an H-bomb—has proven much easier than controlling it: the quest for controlled fusion power is still struggling. Likewise, containing an intelligence explosion might be harder than creating it.

In regard to all these speculations, we don't know where the boundary lies between what may happen and what will remain science fiction. But it's crucial that we explore this issue—one that I have previously addressed on numerous occasions. Environmental degradation, extreme climate change, or unintended consequences of bio-, cyber- and AI technology could trigger serious, even catastrophic, setbacks. We may have a bumpy ride through this century. We've no grounds for assuming that human-induced threats worse than those on our current risk register are improbable: they are newly emergent, so we have a limited time base for exposure to them and can't be sanguine about the ability to cope if disaster strikes. Moreover, in our interconnected world, the consequences would cascade globally.

It is crucial to focus more attention on these x-risks, and that is why this book is so timely. Phil Torres gives a comprehensive survey of the possible risks that have been discussed. He offers a clear (but scary!) review of the technologies. He also notes that the risk level depends on the number of humans who have the motivation to generate global terror—and, more mundanely, on the vulnerability of ever more complex systems to breakdown as well as innocent error.

There are already established research groups and government bodies addressing more "routine" risks—indeed, most organizations are required to produce a "risk register." But these extreme high consequence/low probability risks, potentially affecting the whole world, have hitherto been seriously addressed by only a small community of serious thinkers, whose ideas are described in the book. There needs to be a much expanded research program, involving

natural and social scientists, to compile a more complete register of possible "x-risks," to firm up where the boundary lies between realistic scenarios and pure science fiction, and to enhance resilience against the more credible ones. The stakes are so high that those involved in this effort will have earned their keep even if they reduce the probability of a catastrophe by a tiny fraction.

Technology brings with it great hopes but also great fears. We mustn't forget an important maxim: the unfamiliar is not the same as the improbable.

This encyclopedic book is especially needed. Let's hope it has a wide resonance—and encourages a more intensive and serious focus on issues on which, it's no exaggeration to say, the fate of future generations depends.

—**Lord Martin Rees**, Astronomer Royal, former president of the Royal Society, member of the Board of Sponsors of the *Bulletin of the Atomic Scientists*, and cofounder of the Centre for the Study of Existential Risk

Preface

The field of existential risk studies can trace its origins back to the end of World War II, when the *Bulletin of the Atomic Scientists* created the Doomsday Clock to represent our collective nearness to a global disaster. Later, the astrobiologist Carl Sagan popularized the Drake equation (section 1.5) in the television series *Cosmos* and published an important commentary on the consequences of a major nuclear conflict.[1] According to Sagan, if humanity survives for the next 10 million years, we could expect some 500 trillion people to come into existence.[2] Thus, an all-out nuclear exchange that causes human extinction would not only kill the entire current human population but close off the possibility of *billions and billions* of future lives ever being lived. This makes extinction scenarios especially worrisome—a class of catastrophes with unique moral significance.[3]

In the mid-1990s, the Canadian philosopher John Leslie published an important book called *The End of the World: The Science and Ethics of Human Extinction*, which covers a wide range of existential risks—although he didn't use that term. Leslie also provided perhaps the most compelling defense to date of the doomsday argument (section 7.1), which implies that we are systematically underestimating the probability of human extinction. The work of Leslie influenced another notable figure, namely, Nick Bostrom, the founding director of the Future of Humanity Institute (FHI) at the University of Oxford. Bostrom's work initially focused on anthropic reasoning, including the observation selection effect (section 1.6), which has some important implications for evaluating the overall risk of annihilation. In 2002, Bostrom published an article in the *Journal of Evolution and Technology* called "Existential Risks: Analyzing Human

Extinction Scenarios and Related Hazards." This formalized the concept of an existential risk, introduced the Maxipok rule (section 1.4), and offered an authoritative outline of the biggest threats to our collective future. Bostrom's 2002 article is largely responsible for the popularity—and publicity—of existential risk studies today, a feat that was helped along by his 2014 best seller *Superintelligence*, which provides a detailed account of the technical and philosophical challenges of creating a "friendly" superintelligence.

Although one could argue that the field hasn't quite reached a "normal science" mode of operation yet—to borrow a term of art from Thomas Kuhn—there is an emerging consensus about the central terms, fundamental concepts, and canonical works of existential risk scholarship.[4] There has also been an explosion of institutes dedicated to (a) studying the various existential risks that haunt our species, and (b) devising strategies to mitigate these risks. Such research organizations include the aforementioned FHI as well as the Centre for the Study of Existential Risk (CSER), Future of Life Institute (FLI), Global Catastrophic Risk Institute (GCRI), and my own X-Risks Institute (XRI). In some cases, high-profile scholars or celebrities have put their weight behind these organizations to increase public awareness. For example, Stephen Hawking, Alan Alda, and Morgan Freeman are all members of FLI's scientific advisory board.

So, the "x-risk ecosystem," as the cofounder of FLI and CSER Jaan Tallinn calls it, has grown into a thriving network of scholars and institutions bridging both popular culture and academia.[5] Yet the field does not so far have a comprehensive "textbook" to guide curious young scholars who would like to make the greatest possible impact on the world.[6] This book—an advanced introduction to existential risks; essentially, a progress report on the field—aims to fill this lacuna, thereby further establishing the field as a legitimate area of intellectual inquiry. It attempts to adumbrate something *resembling* a "paradigm" by integrating a wide range of ideas that bear on the topic.

(See the postscript for discussion.)

The target audience includes undergraduate and graduate students in fields as diverse as philosophy and ethics, political science, engineering, computer science, cognitive science, psychology, terrorism studies, sociology, cosmology, and risk analysis.[7] In addition, policymakers, politicians, entrepreneurs, and other culture shapers should find this book full of timely and useful insight.[8] More than anything, I would like *Morality, Foresight, and Human Flourishing* to inspire bright minds around the globe to think more, and more carefully, about the possible, probable, and preferable futures of our species on this planet—and beyond.[9]

Chapter 1: An Emerging Field

1.1 A Unique Moment in History

One can make a very strong case that humanity has never lived in more peaceful times. According to the Harvard polymath Steven Pinker, violence has been declining since humanity struggled as hunter-gatherers in the Paleolithic, roughly 12,000 years ago. This trend has continued through the twentieth and into the twenty-first century, despite the two world wars, Korean War, Vietnam War, Second Congo War (also known as the African World War), and rise of global terrorism, associated most notably with al-Qaeda, Boko Haram, and the Islamic State. We find ourselves in the midst of (a) what historians call the "Long Peace," a period that began at the end of World War II and during which no two superpowers have gone to war, and (b) what Pinker tentatively dubs the "New Peace," which refers to "organized conflicts of all kinds—civil wars, genocides, repression by autocratic governments, and terrorist attacks—[having] declined throughout the world" since the Cold War concluded in 1989.[1] If you could choose when you would like to live in human history since our debut in East Africa some 200,000 years ago, the most reasonable answer would be, "Today, at the dawn of the twenty-first century. No question!"[2]

But there is a countervailing trend that tempers the good news presented by Pinker's historical analyses: we might also live in the most dangerous period of human history, ever.[3] The fact is that our species is haunted by a *growing swarm of risks* that could either trip us into the eternal grave of extinction or irreversibly catapult us back into the Stone Age. Just consider that humanity has stood in the

flickering shadows of a nuclear holocaust since 1945, when the United States dropped two nuclear bombs on the Japanese archipelago. In the years since this epoch-defining event, scientists have confirmed that climate change and global biodiversity loss are urgent threats with existential implications, while risk experts have become increasingly worried about the possibility of malicious individuals creating designer pathogens that could initiate a worldwide pandemic. Looking further along the threat horizon, there appears to be a number of unprecedented dangers associated with molecular nanotechnology and artificial intelligence.[4] Thus, one only needs simple arithmetic to see that the total number of *existential risk scenarios* has increased significantly since the Atomic Age began, and it looks as if this trend will continue at least into the coming decades, if not further.[5]

Considerations of these phenomena have led some scholars to offer unsettlingly high estimates that a global disaster will occur in the foreseeable future.[6] For example, the philosopher John Leslie argues that we have a 30 percent chance of extinction in the next five centuries.[7] Even more ominously, an "informal" 2008 survey of experts at a conference hosted by the Future of Humanity Institute gave a 19 percent chance of extinction before 2100.[8] And the cosmologist Martin Rees writes in a 2003 book that civilization has a 50-50 chance of surviving the present century.[9] To put this in perspective, consider that the average American has a 1-in-9,737 lifetime chance of dying in an "air and space transport accident."[10] It follows that according to the FHI survey, the average American is at least *1,500 times more likely* to perish in a human extinction catastrophe than a plane crash. Using Rees's estimate, the average American is nearly *4,000 times more likely* to encounter a civilizational collapse than to die in an aviation mishap.[11]

If this sounds unbelievable—and no doubt it does, and should—reflect on how many people would be affected by such a disaster. An

analogous case involves asteroids (see section 2.4). According to statisticians, the average person is more likely to die from an asteroid impact than a bolt of lightning (which itself is more likely to kill the average American than a terrorist attack). In fact, the U.S. National Research Council reports that we should *expect* an average of 91 deaths each year from asteroids striking Earth, even though the *actual* number is almost always zero.12 They calculate this number by considering how many asteroids there are near Earth, how big these asteroids are, and how devastating an impact would be. Averaging the total expected deaths over millennia, they get the counterintuitive results above.[13] So, the comparisons of the previous paragraph might not be that far off the mark: a child born today may have a very good chance of living to see global society destroy itself.[14]

Finally, consider the Doomsday Clock, a metaphor that represents our collective nearness to doom, or midnight. This clock was created in 1947 by the *Bulletin of the Atomic Scientists*, an organization founded by physicists who had previously worked on the Manhattan Project, which built the first atomic bombs. Over time, the minute hand of the clock has moved back-and-forth to track the vicissitudes of world affairs: beginning at 7 minutes to midnight in 1947, it moved to only 2 minutes in 1953 (after the United States and Soviet Union both detonated hydrogen bombs) and then drifted away from doom to 17 minutes before midnight when the Cold War "officially" ended in 1991.[15]

While the Bulletin was originally founded to monitor the dangers posed by the world's nuclear arsenals, it announced in 2007 that "climate change also presents a dire challenge to humanity." Consequently, the clock's minute hand inched from 7 to 5 minutes to midnight. After wavering between 5 and 6 minutes, it moved forward again in 2015 due to "unchecked climate change, global nuclear weapons modernizations, and outsized nuclear weapons arsenals," which "pose extraordinary and undeniable threats to the continued existence of humanity." A year later, the Bulletin decided to keep the

clock set at 3 minutes to midnight, writing that "the world situation remains highly threatening to humanity, and decisive action to reduce the danger posed by nuclear weapons and climate change is urgently required."[16]

But 2017 saw the minute hand tick 30 seconds closer to doom, reaching the highest level of danger since 1953. This was largely due to two factors, both enabled by what one could describe as a *zeitgeist of anti-intellectualism* that currently pervades Western, especially American, political culture. As the Bulletin's official statement puts it, an

> *already-threatening world situation was the backdrop for a rise in strident nationalism worldwide in 2016, including in a U.S. presidential campaign during which the eventual victor, Donald Trump, made disturbing comments about the use and proliferation of nuclear weapons and expressed disbelief in the overwhelming scientific consensus on climate change.*[17]

On the same day of this announcement, the cosmologist Lawrence Krauss and international affairs expert David Titley, both of whom help maintain the Doomsday Clock, published a *New York Times* op-ed titled "Thanks to Trump, the Doomsday Clock Advances toward Midnight." In their words,

> *The United States now has a president who has promised to impede progress on both [curbing nuclear proliferation and solving climate change]. Never before has the Bulletin decided to advance the clock largely because of the statements of a single person. But when that person is the new president of the United States, his words matter.*[18]

The point is that *many* leading experts believe the threat of an existential catastrophe to be significant.[19] Before 1945, overseeing a

Doomsday Clock would have been utterly nonsensical, since the existential threats posed by nature are relatively improbable (see below). Yet today, the clock stands at two-and-a-half minutes before midnight, and it appears poised to tick forward again in 2018. To be sure, the predicament of *Homo sapiens* on Earth has always been precarious—consider that we are the *only remaining species* of *Homo* on the planet, our relatives the Neanderthals having died out about 40,000 years ago—but changes to the global climate and ecosystem along with the development of powerful new technologies are making our continued survival more uncertain than ever.

1.2 What Are Existential Risks?

The concept of an **existential risk** (ER) was formalized by the Oxford philosopher Nick Bostrom in a 2002 paper.[20] To understand this term's definition, it is helpful to know that Bostrom is a prominent figure within the *transhumanist movement*. According to transhumanism, person-engineering technologies will enable us, if we wish, to modify aspects of our bodies and brains, perhaps resulting in a new species of *posthumans*, while world-engineering technologies will enable us to radically redesign the environments in which we live to make them more conducive to flourishing (where some of these environments could be simulated rather than "real").[21] Whereas bioconservatives embrace "therapeutic" but not "enhancive" interventions on the human organism, transhumanists advocate exploring what could be a vast space of posthuman modes of being, some of which may be *far better* in certain moral respects than our current human mode.[22] Thus, transhumanism has both descriptive and normative components.[23] (See Box 1.)

To be clear, most transhumanists are careful to emphasize that "can" does not imply "ought"—that is, just because we are able to modify our phenotypes doesn't mean that we are obliged to do so. Rather, humanity should proceed according to something like the

"precautionary principle," which states that "an action should not be taken if the consequences are uncertain and potentially dangerous,"[24] or perhaps the philosopher Max More's "proactionary principle," which argues that

> *People's freedom to innovate technologically is highly valuable, even critical, to humanity. This implies several imperatives when restrictive measures are proposed: Assess risks and opportunities according to available science, not popular perception. Account for both the costs of the restrictions themselves, and those of opportunities foregone. Favor measures that are proportionate to the probability and magnitude of impacts, and that have a high expectation value. Protect people's freedom to experiment, innovate, and progress.*[25]

Box 1. As the AI entrepreneur Riva-Melissa Tez puts it, transhumanism "sounds weirder than it actually is."* It is simply the idea that, within certain ethical boundaries and guided by the epistemic value of "philosophical fallibilism," we should not be afraid to use technology to improve the human condition, which is currently marked by widespread suffering, the hedonic treadmill, disease, senescence, and death. There are a couple of points worth noting here: First, we have already vastly improved our situation through the use of technologies, some of which—such as clothes, glasses, telescopes, prosthetics, psychoactive pharmaceuticals, pacemakers, cochlear implants, smartphones, and the Internet—directly alter, extend, and enhance our phenotypes. Compared to our Paleolithic progenitors, most modern humans are "transhumans" already—virtually a different species. Second, humanity is evolving anyway due to ongoing mechanisms like natural selection and genetic drift, and indeed some scientists believe that human evolution has actually accelerated in recent centuries. Thus, we will someday become "posthumans" even if

bioconservative policies are universally implemented, just as some of our ancient Hominini relatives became "post-Australopithecines" by evolving into *Homo sapiens*. Since biological evolution is a non-teleological process—meaning that *every state* is an in-between state; there is no finalistic "resting place" at which all human genetic changes cease[†]—why not try to take control of our own evolution through intentional cyborgization, to direct our lineage toward future states marked by improved health, happiness, longevity, intelligence, morality, and so on? This isn't such a radical idea after all—and in fact one could argue that it is the default, albeit tacit, view of many Westerners today. It is certainly the direction in which our technological civilization appears to be headed.

* ogilvy do. 2015. Technology: Making the World a Better Place. YouTube. URL: https://www.youtube.com/watch?v=i5t1BQUbSB4&t=43s.

† Or, as Charles Darwin put it, "not one living species will transmit its unaltered likeness to a distant futurity." Thus, there is a sense in which bioconservatism is a *nonstarter*. See Darwin, Charles. 2007. *On the Origin of Species: By Means of Natural Selection or The Preservation of Favored Races in the Struggle for Life*. New York, NY: Cosimo Classics.

Having outlined the basics of transhumanism, we can now make sense of Bostrom's definition of an existential risk:

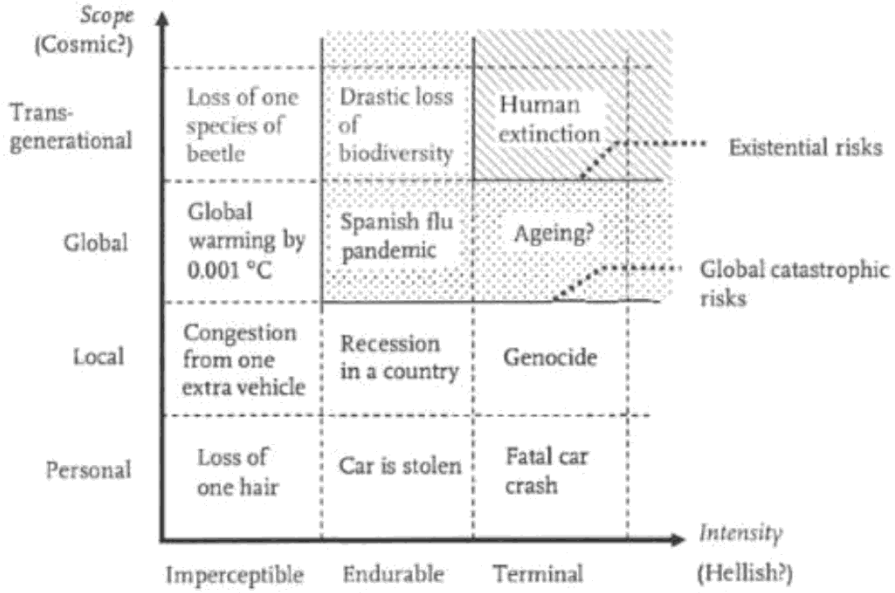
> *An existential risk is one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development.*[26]

Thus, there are two general categories of existential risk scenarios: (i) total annihilation, and (ii) an irreversible curtailing of our potential. The first disjunct is straightforward: the lineage of Earth-originating intelligent life terminates. This outcome is *binary*: we

either live or die, persist or desist, remain extant or go extinct. The second disjunct is not so clear-cut, given the normativity of "desirable." It is here that transhumanism enters the axiological picture. From this perspective, the ultimate goal of civilization is to safely reach a state of *technological maturity*, which Bostrom limns as "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved."[27] It follows that a catastrophe—in this case, an endurable catastrophe of type (ii)—counts as "existential" if and only if it prevents our species from realizing the posthuman promise of "mature technology."

In addition to a definition of "existential risk," Bostrom offers three typologies of risks in general.[28] These are based on a conceptual decomposition according to which a risk equals *the probability of an event multiplied by its consequences*. (Note that this entails that a high-consequence risk could be significant even if it is extremely improbable.) With respect to the first variable, there are multiple interpretations of probability, such as the propensity, frequency, and Bayesian interpretations, none of which we will here explore. With respect to the second, Bostrom analyzes the consequences of an event into two subcomponents: *scope* and *intensity*. Scope refers to how many people are affected, and intensity to how bad the effects are. The result is a two-dimensional typology, Figure A, in which existential risks occupy the top right box of transgenerational-terminal events (where "terminal" is stipulated to include some endurable events).[29]

Figure A. Two-Dimensional Typology of Risks



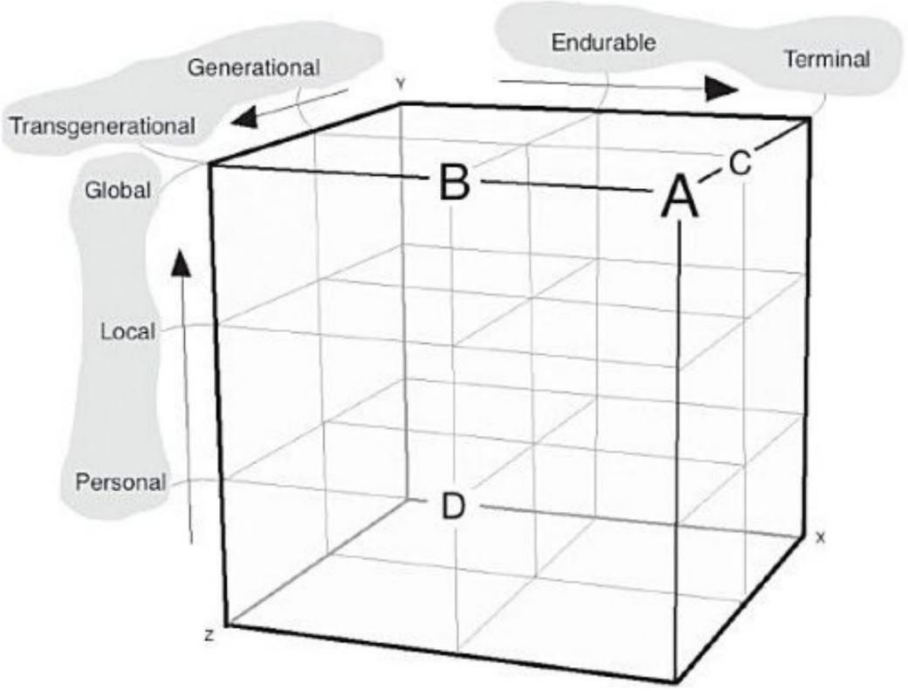
Source: Nick Bostrom and Milan Ćirković. 2008. Introduction. In Nick Bostrom and Milan Ćirković (editors), *Global Catastrophic Risks*. Oxford: Oxford University Press.

But we can refine Bostrom and (his coauthor) Milan Ćirković's typology by further decomposing the scope of a risk's consequences into *spatial* and *temporal* sub-subcomponents. This is motivated by the truism that risks can have a range of different spatiotemporal ramifications. For example, a germline mutation could have limited consequences within a population, yet these consequences could linger for an indefinite number of future generations. (Where would this risk fit in Figure A?) Similarly, a catastrophe could *instantaneously* kill 1 billion people at a given timeslice or *incrementally* kill the same number over the course of a century. Distinguishing between these scenarios is important because our responses to each might require quite divergent counterstrategies. Thus, insofar as Bostrom and Ćirković's typology is intended to provide an exhaustive classification of risks, it appears inadequate.[30]

By adopting a decomposition of risks according to the three

properties of intensity, spatial scope, and temporal scope, one gets the three-dimensional typology of Figure B. In this figure, existential risks occupy two positions: (1) the node marked "A," which corresponds to global-*terminal*-transgenerational catastrophes, and (2) the node marked "B," which encompasses those global-*endurable*-transgenerational events that, by definition, prevent humanity from ever attaining technological maturity. Furthermore, germline mutations correspond to node D, while aging (which fits uncomfortably in Figure A, as indicated by the question mark) corresponds to node C—that is, it affects everyone globally with death but doesn't entail our extinction.[31]

Figure B. Alternative Typology of Risks Based on the Properties of Intensity (x-axis), Spatial Scope (y-axis), and Temporal Scope (z-axis)



Note: Consequences get worse as one follows the arrows.

Whichever typology one finds most useful, the key idea is that existential risks constitute *worst-case scenarios* for humanity—resulting in what the philosophers Ingmar Persson and Julian Savulescu call "Ultimate Harm"—given our potential to reach new and better modes of being.[32]

Other important features of existential risks are the following:

(a) They are *singular events* that can only happen once in a species' lifetime; this makes them quite unique among all the other types of risks that we face. For example, we can talk about a human extinction event happening tomorrow but not about it having happened yesterday; and while we can talk about an endurable existential catastrophe having happened yesterday, we would not be able to do anything about it, to reverse the outcome. If an existential risk were to occur, *the game would be over and humanity would have lost*.

(b) Since existential risks have the properties of (a), our strategies for avoiding them must rely entirely on *anticipation* rather than *retrospection*. As Bostrom writes, "The reactive approach—see what happens, limit damages, and learn from experience—is unworkable. Rather, we must take a proactive approach."[33] This means that humanity must employ "unnatural" modes of thinking, since our typical way of avoiding bad future circumstances is to update our world models in response to past mistakes made by ourselves or others. But there is no possibility of learning from the mistakes that humanity made leading up to an existential catastrophe so that we don't encounter another existential catastrophe later on.[34]

(c) These points suggest that individuals and governments are unlikely to make existential risk reduction a top priority. Since an effective risk mitigation program would result in the *absence* rather than *presence* of an observable event, a record of success

could lead to complacency, causing people to question whether money is being well-spent. The risk analyst Nassim Taleb makes this point in the context of "black swans," or game-changing incidents that are inadequately expected:

> *It is difficult to motivate people in the prevention of Black Swans…. Prevention is not easily perceived, measured, or rewarded; it is generally a silent and thankless activity. Just consider that a costly measure is taken to stave off such an event. One can easily compute the costs while the results are hard to determine. How can one tell its effectiveness, whether the measure was successful or if it just coincided with no particular accident? … Job performance assessments in these matters are not just tricky, but may be biased in favor of the observed "acts of heroism." History books do not account for heroic preventive measures.*[35]

(d) Even more, the reduction of existential risks constitutes a *global public good*, meaning that it is both *non-excludable* (i.e., it is not possible to prevent those who haven't paid for this service from benefiting) and *non-rivalrous* (i.e., it is not the case that one person benefiting prevents others from benefiting). This is notable because markets don't typically provide such goods, since producers can only retrieve a small amount of value relative to the costs of production. As Bostrom elaborates this point,

> *In fact, the situation is worse than is the case with many other global public goods in that existential risk reduction is a strongly* transgenerational *… public good: even a world state may capture only a small fraction of the benefits—those accruing to currently existing people. The quadrillions of happy people who may come to exist in the future if we avoid existential catastrophe would be willing to pay the present generation astronomical sums in return for a slight*

increase in our efforts to preserve humanity's future, but the mutually beneficial trade is unfortunately prevented by the obvious transaction difficulties.[36]

So, existential risks form a special class of catastrophes that pose genuinely unique challenges to civilization.

Before moving on to the next section, we should consider a related topic of interest, namely, **global catastrophic risks** (GCRs). Bostrom and Ćirković define GCRs "loosely" as events "that might have the potential to inflict serious damage to human well-being on a global scale." They suggest that a disaster causing "10 million fatalities or 10 trillion dollars worth of economic loss … would count as a global catastrophe, even if some region of the world escaped unscathed."[37] Other scholars have defined GCRs as events that result in one-fourth of the human population dying, or "threats that can eliminate at least 10% of the global population."[38] In Figure A, GCRs encompass risks within the light and dark gray boxes—meaning that existential risks are a special case of global catastrophic risk. With respect to Figure B, we can define GCRs as any risk that (a) has the property of *being global*—that is, it instantiates a node on the top level of the diagram—and (b) causes sufficiently severe harm to human civilization.[39]

Given that the probability of a risk tends to *increase* as its consequences *decrease*, the chance that one or more GCRs occur this century should exceed the probabilities assigned to an existential catastrophe occurring—which, once again, range from 19 percent to 50 percent.[40] More concretely, a pandemic that kills 1 billion people will be more probable than one that causes human extinction; the same goes for an asteroid impact, nuclear war, nanotech accident, and so on. Thus, if we believe that human extinction from a pandemic has, say, a 1 percent chance of happening per decade, we should believe that 1 billion people dying in a pandemic has a *greater than or equal to 1 percent chance* of happening over the same period.41 In general, the smaller the consequences, the higher the probability.

Furthermore, insofar as the *timing* of non-existential GCRs is random—which is not an implausible assumption, since (a) many natural risks are in some sense "random," and (b) studies have actually shown that "the onsets of wars [are] randomly timed"—we should (weakly) expect them to cluster together in time.[42] For example, if there is a constant probability of 0.05 that a GCR will occur per decade, and if a GCR occurs during the first decade of a new century, the probability of a GCR occurring the second decade will actually be *higher* than one occurring the third decade, or any decade afterwards. The reason is that for a GCR to occur next during the third decade, it would have to *not have occurred* during the second. Thus, *two conditions* must hold: (i) no GCRs during the second decade, and (ii) a GCR during the third decade. To calculate the probability of this joint state of affairs, one multiplies the probability of (i), or 0.95 (from 1 minus 0.05), by the probability of (ii), or 0.05. This yields a probability of 0.0475 for a GCR happening in the third decade, which is, of course, lower than the 0.05 probability of a GCR happening in the second decade. As the mathematician William Feller once put it, "To the untrained eye, randomness appears as regularity or tendency to cluster," meaning that we should not be *too* surprised if a series of global catastrophes unfolds one after another.[43]

While this book focuses primarily on existential risks, given their unique moral status (see section 1.4), GCR issues will nonetheless appear throughout.

1.3 Types of Existential Risks

There are different ways to taxonomize existential risks depending upon one's theoretical or practical goals. In a 2013 paper, Bostrom offers a four-part scheme that includes human extinction, permanent stagnation, flawed realization, and subsequent ruination. With respect to Figure B, the first is a global-terminal-transgenerational disaster (node A), whereas the latter three are global-endurable-

transgenerational disasters (node B). Taking these in turn:

(i) **Human extinction**. About 99.9 percent of all species that have ever existed on Earth have gone extinct, and the average mammal survives for only about 2.5 million years.[44] As Carl Sagan put it, "Extinction is the rule. Survival is the exception."[45] Here we should expand the semantics of "human" to include not just *Homo sapiens* but Earth-originating intelligent life in general, independent of its material substrate (e.g., living cells or microchips). This is important because if the cyborgization trend of integrating biology and technology, organism and artifact, continues, our descendants could become sufficiently different from us to constitute a new species: *Homo cyborgensis*, or something of the sort.[46] If a future posthuman population of *Homo cyborgensis* were completely decimated, we should like this to count as an existential catastrophe too.

(ii) **Permanent stagnation**. This scenario would occur if (i) does not obtain yet humanity never reaches a state of technological maturity. Bostrom distinguishes several types of stagnation, including (a) *unrecovered collapse*, where "much of our current economic and technological capabilities are lost and never recovered," (b) *plateauing*, where "progress flattens out at a level perhaps somewhat higher than the present level but far below technological maturity," and (c) *recurrent collapse*, which would entail "a never-ending cycle of collapse followed by recovery."[47] To this taxonomy we can add a "catch-all" category that includes any combination of these scenarios, such as long plateaus punctuated by collapse, followed by recovery to another plateau, followed by unrecovered collapse.

(iii) **Flawed realization**. This involves reaching "technological maturity in a way that is dismally and irremediably flawed." In other words, we achieve a posthuman state that realizes only "a

small part of the value that could otherwise have been realized."[48] Bostrom identifies two instances of this outcome. The first, *unconsummated realization*, occurs when future technologies fail to achieve states of high value. For example, it could be the case that future artificial intelligences (AIs) inherit the world, but that these AIs do not have conscious experiences like we do. As the philosopher Susan Schneider rightly emphasizes, a world full of unconscious machines—even if these machines were to build a complex, advanced civilization throughout the known universe—would be far less valuable than one in which even a single conscious being exists.[49] The result would be an existential catastrophe.

The second type of flawed realization is *ephemeral realization*. This results when "humanity develops mature technology that is initially put to good use. But the technological maturity is attained in such a way that the initial excellent state is unsustainable and is doomed to degenerate." For example, it could be that achieving technological maturity leads to significant social, political, or cultural divisions that over time cause major conflicts to break out, and that these conflicts bring about an extinction or permanent stagnation disaster. As Bostrom puts it, "There is a flash of value, followed by perpetual dusk or darkness."[50]

Box 2. Of all the existential risk categories here enumerated, extinction appears to be the least likely. The reason pertains to what might be called the *last few people problem*: one can readily devise hypothetical narratives in which a large number of humans perish, but it is rather hard to envision how the last people on the planet follow their conspecifics to the grave. This problem emerged from a 2009 special issue of the journal Futures, co-edited by Bruce Tonn and Donald MacGregor, in which scholars were tasked with concocting extinction scenarios. As Tonn and MacGregor write, "It is quite easy to imagine events that could

lead to a rapid and massive loss of human life…. [But most] of the scenario writers found that indeed it was difficult to kill off the last few humans and most were surprised … for this to be the case. We speculate that is the good news coming out of this special issue."*

* Tonn, Bruce, and Donald MacGregor. 2009. Are We Doomed? *Futures*. 41(10): 673–675.

(iv) **Subsequent ruination**. Our final category occurs when (i) through (iii) fail to obtain, meaning that we reach an unflawed state of technological maturity. Our species appears to have accomplished the ultimate triumph. Yet *further developments* in technology, social institutions, government, and so on bring about either the termination of our lineage or an irreversible decline in our quality of life.[51] (See Box 2.)

While this taxonomy is helpful for understanding different features of possible worst-case futures, we will adopt a different approach that focuses not on the *outcomes* of various scenarios but on those scenarios' *causes*. We can call this the **etiological approach**. Attending to the underlying causes of different scenarios is arguably more important because when one understands the causes behind an effect, one can avoid the effect by intervening on the causes. For example, if you know that a brake failure was the cause of your car racing through a red light, then you can prevent future traffic violations by fixing the brakes. Similarly, if you know that smoking causes lung cancer, then you can reduce your chances of a bad oncological diagnosis by refraining from smoking. Thus, specifying the etiology of different outcomes is crucial for avoiding a catastrophe.

The broadest causal distinction is between **natural risks** and **anthropogenic risks**. Supervolcanic eruptions, natural pandemics, and asteroid or comet impacts are the most worrisome natural risks. Less concerning are supernovae, gamma-ray bursts, galactic center outbursts, superstrong solar flares, and black hole mergers or

explosions. The universe could also contain any number of currently unknown risks to our survival. Perhaps a discovery by physicists 50 years from now will reveal a new type of natural danger that is as unimaginable to twenty-first-century humans as the threat of gamma-ray bursts was to those in the Pleistocene. Or it could be that no possible future science can reveal certain threats because understanding them requires a different *kind of mind* than what natural selection gave us. As far as contemporary science is concerned, though, the overall probability of a natural existential risk destroying humanity per century is almost certainly less than 1 percent, and arguably *far less*.[52]

Moving on to the category of anthropogenic risks, this contains a diverse range of distinct and overlapping phenomena. The most significant subtype stems from what we will refer to as **agent-tool couplings**.[53] We can define an agent somewhat crudely as any entity capable of making its own decisions in pursuance of its own goals, whatever they happen to be. There are many *degrees* of agency in the world: for example, a heat-seeking missile has a certain degree of agency since it can navigate space-time in response to inputs relating to its target. The agents most relevant in this context, though, are those with general intellectual abilities, whether human or machinic in nature, such as apocalyptic terrorists and artificial superintelligence. As for the tool half of the coupling, this includes any advanced technology with the capacity to cause an existential catastrophe. We can call these **weapons of total destruction** (WTDs), on the model of "weapons of mass destruction" (WMDs). Such technologies could be actual (e.g., nuclear weapons) or merely anticipated (e.g., molecular nanotechnology), and indeed many existential risk scholars believe that future anticipated technologies will likely pose far greater risks than those around today. There could also be technologies that are not currently anticipated by anyone but that will introduce novel hazards for humanity.

The "agent-tool" concept is essential for existential risk studies

because, bracketing the possibility of malfunction, dangerous technologies require a suitable agent to *use* them to cause harm. It follows that to assess the relevant risks, one must evaluate *both* the artifacts *and* their users. This framework also emphasizes that there are two definite variables—the agents and the tools—that could be intervened upon to reduce overall existential risk. Thus, chapter 6 explores "tool-oriented" and "agent-oriented" strategies for reducing existential risks.

Another subtype of anthropogenic risk derives from **unintended consequences**. The most troubling unintended consequences today are climate change and biodiversity loss, although there are also potential risks associated with physics experiments and geoengineering. As all responsible citizens of the world should know, climate change is the result of greenhouse gas emissions, which are a byproduct of burning fossil fuels. This is arguably the first unintended consequence in human history with genuinely existential implications—but it will probably not be the last. Indeed, when automobiles with internal combustion engines were adopted *en masse* in the early twentieth century, they were widely praised as a solution for urban pollution, a major health problem at the time, which took the form of horse excrement and carcasses. (This also resulted in the spread of illness by the "disease vectors" of flies.[54]) The unfortunate irony is that automobiles have become one of the greatest contributors to a global-scale calamity that threatens the future stability of civilization itself. While climate change is a primary cause of biodiversity loss, which has initiated a new mass extinction, biodiversity loss can also exacerbate climate change—for example, through the elimination of carbon sinks, which remove carbon dioxide from the atmosphere.

As for physics disasters: while this scenario appears highly improbable on our best current theories, these theories could be flawed. A high-powered particle accelerator could thus accidentally initiate a catastrophe with planetary or even *cosmic* consequences. Geoengineering, which involves redesigning one or more physical

features of our planetary spaceship (i.e., Earth), poses several perils. For instance, a group or government could unilaterally opt to inject particles into the stratosphere to block incoming sunlight, thereby reducing the negative consequences of "too much" atmospheric carbon dioxide. Although this could, it appears, save humanity in a climate emergency (see section 6.5), it could also have severe unintended repercussions. Alternatively, if the injection of particles into the stratosphere were to work but then suddenly stop for some reason, surface temperatures could rebound too quickly for civilization to adapt.

Finally, we will examine a range of risks that don't directly arise from either agent-tool couplings or the unintended consequences of human activity. This motley group includes:

(a) *Simulation shutdown*. However dubious this may initially sound—and it should sound dubious to any good skeptic—there are some rather compelling, albeit esoteric, reasons for believing that we might live in a computer simulation. If so, this would introduce the possibility that our simulation gets shut down, thereby resulting in an existential catastrophe.[55]

(b) *Bad governance*. Unwise governments could ignore the established science behind climate change and biodiversity loss—and, indeed, many governments are doing precisely this. They could also engage in arms races involving molecular nanotechnology or superintelligence, both of which would likely yield "winner-take-all" situations. If such a race were to occur and if the "winner" were to "take all," humanity could find itself under the control of a totalitarian state—one that might stifle further technological development, not to mention human happiness.

(c) *Something completely unforeseen*. It would be imprudent to believe that we—apes with big foreheads—know all the risks to our species. There could be unknown natural risks, unanticipated

future technologies, new types of dangerous agents, and unintended consequences from, say, colonizing space. A book like this written 200 years from now could contain 3 (or 20) times as many chapters focusing on scenarios of which we haven't the slightest inkling. Indeed, the existential risks explored throughout this manuscript could be relegated to the appendix, being seen as the *least worrisome* relative to the new, futuristic threat environment of our descendants.

* * *

There are a few conceptual distinctions worth mentioning before moving on. First, consider the difference between **state risks** and **step risks**.[56] The former arise from being in a certain state, whereas the latter arise from transitioning between states. To illustrate, dying in a car accident is a state risk: the danger is associated with a specific situation, namely, driving a car, and the longer that one is in this situation, the greater the risk. Many risks from nature are state risks. In contrast, walking onto a train from the railway platform constitutes a step risk: the danger is associated with the transition from being on the platform to being in the train. Thus, in the London Underground one hears the warning, "Mind the gap." Once inside the train, the danger is gone (although one then encounters a new state risk). The existential danger posed by superintelligence may be a step risk.

There are also what we might call **context risks**. These are big-picture phenomena that *frame* our existential predicament on the planet. The most notable context risks are climate change and biodiversity loss. Such risks have implications for the overall probability of doom, even if they are themselves unlikely to bring about an existential catastrophe (that is, as a proximate cause of the disaster). Put differently, contexts risks can *modulate* the dangers posed by other risk scenarios. A simple intuition pump illustration is the following: imagine two worlds, A and B. World A finds itself beset by

social turmoil, economic meltdowns, and political strife as a result of environmental atrophy, whereas the climate and biosphere of world B remain in relative homeostasis. Now imagine that both worlds contain 10,000 nuclear weapons. In which world is nuclear conflict more likely to break out *a priori*? The answer is, obviously, world A. The capacity for conflict-multiplying context risks to raise or lower the tide of all other existential threats makes phenomena of this sort especially important to prioritize. (This is a crucial point that I hope readers will dwell on.)

1.4 Why Care about Existential Risks?

Nothing is too wonderful to be true, if it be consistent with the laws of nature.
—Michael Faraday

The global population today is 7.5 billion. Let's say that a pandemic spreads across Europe, killing 100 million people. How bad would this be? Most would agree that it would be quite devastating. Now let's say that 100 million more people die from the disease. How bad would *this* be? It seems like this second wave of deaths would be just as bad as the first: 200 million people dying is twice as horrible as 100 million people dying. Now imagine this continuing 74 times (where 74 x 100 million = 7.4 billion), with each instance of 100 million deaths being an equivalently bad moral tragedy. The global population would then be only 100 million people. Again, we can ask: If this last group were to die from the pandemic, how bad would it be? Would it be just as bad as each past instance of 100 million people dying—or might it be *worse*?

The philosopher Derek Parfit, echoing Sagan's idea discussed in the preface of this book, argues that the last 100 million people dying would not only be worse than all the other instances of 100 million people dying, but *profoundly worse*. The reason is that, as Parfit writes,

> *Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between [nearly all and actually all people dying] may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second.*[57]

We can add to Parfit's thesis an alternative scenario, given the second disjunct of our definition of existential risks: consider a world in which there are no incidents of mass dying but some unfortunate event causes civilization to sink into a permanent state of technological deprivation. The result would be that we fail to reach technological maturity and exploit our *cosmic endowment of negentropy* (where "negentropy" is a portmanteau of "negative entropy," i.e., the stuff that enables living systems to create and maintain order in the universe).[58] From the transhumanist point of view, the result would be, all things considered, no less tragic than if humanity were to go extinct.[59]

A key idea here is that the potential *value* of our posthuman future could be unimaginably huge. For example, one estimate suggests that a total of "a hundred thousand billion billion billion"—that is, a 1 followed by 32 zeros, or 100,000,000,000,000,000,000,000,000,000, 000—humans could someday populate the universe.[60] These people might colonize a large fraction of our future light cone, use enhancement technologies to radically augment their cognitive and moral capacities, live indefinitely long lives through rejuvenation therapies, upload their minds to achieve a kind of digital immortality, and perhaps even convert entire planets into supercomputers that run simulations in which conscious beings live happy, worthwhile lives (thereby increasing the total amount of well-being in the cosmos, which some ethical theories prescribe).[61] As Parfit puts the point, "Life can be wonderful as well as terrible, and we shall increasingly

have the power to make life good. Since human history may be only just beginning, we can expect that future humans, or supra-humans, may achieve some great goods that *we cannot now even imagine*."[62] In a phrase, the expected value of the future is *astronomically high* given the potential number and nature of our posthuman descendants. Let's call this the **astronomical value thesis**.[63]

This leads Bostrom to argue that "the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole." In other words, we should behave according to the following "rule of thumb for such impersonal moral action," dubbed **Maxipok**:

> *Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe.*[64]

One can think of our predicament as follows: the present moment —a century that the Long Now Foundation writes as "02000" to encourage "deep time" thinking—is a narrow foundation upon which an extremely tall skyscraper rests.[65] The entire future of humanity resides in this skyscraper, towering above us, stretching far beyond the clouds. If this foundation were to fail, the whole building would come crashing to the ground. Since this would be astronomically bad according to the above thesis, it behooves us to do everything possible to ensure that the foundation remains intact. The future depends crucially on the decisions we make today, just as the present depends on the decisions made by our ancestors, and this is a moral burden that everyone should feel pressing down upon their shoulders.[66]

While one might accept that every human perishing tomorrow would be an unthinkable catastrophe, one might also object that there is no particular reason to value the lives of people who do not yet exist. Why should current people care about generations that are born

100, 10,000, or even 100 million years from today? What obligations do we really have to future people in some far-off, exotic futureland? Many moral philosophers respond that *when* one exists should be irrelevant to that person's moral status. By analogy, *where* one exists should be—it appears correct to assert—irrelevant to whether or not one matters ethically: e.g., the suffering of a child in Johannesburg is just as bad as the suffering of a child in Copenhagen, Beijing, or Honolulu. And since modern physics reveals that space and time form a unified four-dimensional continuum (called "spacetime"), there don't appear to be any *fundamental* reasons for privileging one dimension over another, meaning that "affecting a temporally distant individual in the future is similar to affecting a spatially distant individual" right now.[67] If one rejects "space discounting" (or devaluing the lives of people who are spatially distant from us), one should also reject "time discounting" (or devaluing the lives of people who are temporally distant from us).

Furthermore, as the risk expert Jason Matheny observes, time discounting future lives yields conclusions that "few of us would accept as being ethical."[68] For example, if one were to discount future "lives at a 5% annual rate, a life today would have greater intrinsic value than a billion lives 400 years hence"—i.e., a single person dying this evening would constitute a *worse moral tragedy* than a global catastrophe that kills 1 billion people in four centuries.[69] Similarly, a 10 percent annual discount rate would entail that one person today is equal in value to an unfathomable 4.96×10^{20} people 500 years from now.[70] This line of reasoning appears to be not only misguided but *outrageously wrongheaded*, from which it follows that discounting human lives is deeply problematic.[71]

The futurist Wendell Bell offers seven additional reasons that contemporary generations have obligations to future generations. These are:

(1) *A concern for present people implies a concern for future people.* There is no "clear demarcation … between one generation and the next," meaning that "a concern for people living now carries us a considerable way into caring about future people." Imagine that you have children who have children. You care about your grandchildren, who will one day care about their own grandchildren. The result is an unbroken chain of caring that extends indefinitely into the future.

(2) *Thought experiments in which choosers do not know to which generation they belong rationally imply a concern for both present and future people.* If one knows nothing about which generation one will live and is asked "to choose how each generation ought to behave, consuming now or saving and preparing for the future," rational choosers will "allow for the well-being of both present *and* future generations." (This thought experiment borrows from John Rawls's idea of the "original position," in which people select principles upon which society will be based without knowing anything about their gender, ethnicity, social status, and so on.[72]) It follows that "we ought to care about the well-being of future people because that is what rational people would choose to do if they did not know what generation they were in."

(3) *Regarding the natural resources of the earth, present generations have no right to use to the point of depletion or to poison what they did not create.* Since natural resources were not produced by any human, "everyone has a right to their use, including members of future generations." Therefore, "the members of the present generation have an obligation to future generations of leaving the earth's life-sustaining capacities in as good a shape as they found them or of providing compensating benefits of life-sustaining worth equal to the damage that they do."

(4) *Past generations left many of the public goods that they created not*

only to the present generation but to future generations as well. This suggests that "no generation has the right to use up, totally consume, or destroy the existing human heritage, whether material, social, or cultural, so that it is no longer available to future generations."

(5) *Humble ignorance ought to lead present generations to act with prudence toward the well-being of future generations.* We are only beginning to understand the universe, and we have only the vaguest sense of "what the human destiny is or might become." Thus, "weighted with such ignorance, the present generation ought to act prudently so as not to threaten the future survival and well-being of the human species."

(6) *There is a prima facie obligation of present generations to ensure that important business is not left unfinished.* The term "important business" here refers to "human accomplishments, especially exceptional ones in science, art, music, literature, and technology, and also human inventions and achievements of organizational arrangements, political, economic, social, and cultural institutions, and moral philosophy." Both this and the previous point clearly connect to the transhumanist goal of reaching new and better modes of being.

(7) *The present generation's caring and sacrificing for future generations benefits not only future generations but also itself.* One way to give life meaning is through engagement and altruistic sacrifice. In other words, "it is through being concerned for other people, both living and as yet unborn, that a person achieves self-enrichment and personal satisfaction." As Bell adds, "Genuinely caring about future generations and taking effective action to benefit their well-being are objective and rational answers to the contemplation of one's own death and the feelings of futility and despair it produces. Thus, we can strengthen ourselves by creating

a community of hope."[73]

So, there are compelling reasons for caring about the well-being of future people and, therefore, allocating a nontrivial sum of resources for existential risk research. From a methodological standpoint, this is why the present book considers a wide range of risk scenarios, including some that have a *prima facie* "sci-fi" flavor: given the astronomically high stakes involved, even risky phenomena that seem, from a "pre-theoretic" perspective, unlikely warrant further investigation.[74] Perhaps future research will reveal certain scenarios to be less problematic than initially expected, in which case we can safely ignore them; but it might also show them to be *worse* than anyone imagined, thus requiring immediate action to curb a cataclysm. The only way to know is to put these ideas—all of them, despite any prior prejudices (see section 1.6)—under the electron microscope of critical analysis and to go from there. As Rees eloquently puts it in the foreword of this book, "The stakes are so high that those involved in this effort will have earned their keep even if they reduce the probability of a catastrophe by a tiny fraction."

1.5 Fermi and Filters

Let's now consider some general features of our place in the universe, beginning with the **Fermi paradox**. Named after the physicist Enrico Fermi, who worked on the Manhattan Project, this paradox originated during a 1950 luncheon conversation about the possibility of other civilizations populating the universe. After pondering the issue, Fermi exclaimed, "Where is everybody?" The reasoning goes like this: some 10 billion galaxies and 1 billion trillion stars exist in the observable universe. A certain percentage of these stars will likely have Earth analogs in the habitable or "Goldilocks" zone, the region around a star where conditions are suitable for liquid water and, therefore, carbon-based lifeforms. Given these facts, we should expect a large number of

technologically advanced civilizations to exist—that is to say, *even if* the probability of an advanced civilization developing on any given exoplanet is minuscule, the sheer number of exoplanets in the cosmos should make advanced civilizations abundant.

Yet, dubious anecdotes and grainy footage aside, we see no legitimate signs of extraterrestrial life crying out for cosmic companionship in the darkness of space. We have encountered no aliens with imperialistic ambitions to dominate the galaxy. We find no rapacious swarms of von Neumann probes buzzing around us—that is, spacecraft capable of mining resources throughout the universe to create copies of themselves, thereby producing an exponential expansion of probes in all three dimensions. And we have detected no verifiable squeaks in the form of nonrandom electromagnetic signals washing up against our planetary island.[75] This is the Fermi paradox: the skies are silent when they should be noisy.

Or, perhaps there is a flaw in the above reasoning. In 1961, the astrophysicist Frank Drake proposed an "equation" that attempts to specify all the crucial variables that scientists must consider to calculate the total number of communicable civilizations in the universe. The result is the **Drake equation**, which states that $N = R^* \times f_p \times n_e \times f_l \times f_i \times f_c \times L$. These variables stand for the following:

N is the total number of communicable civilizations.

R^* is the rate of formation of stars suitable for the development of intelligent life.

f_p is the fraction of those stars with planetary systems ("p" for planets).

n_e is the number of planets, per solar system, with an environment suitable for life ("e" for ecologically suitable).

f_l is the fraction of suitable planets on which life actually appears ("l" for life).

f_i is the fraction of life bearing planets on which intelligent life emerges ("i" for intelligence).

f_c is the fraction of civilizations that develop a technology that releases detectable signs of their existence into space ("c" for communicative).

And L is the length of time such civilizations release detectable signals into space.[76]

It is difficult to determine accurate values for each of these variables, and consequently estimates have varied dramatically. For example, Drake and others initially calculated that the number of civilizations in the Milky Way could range between 1,000 and 100 million. In *Cosmos*, Carl Sagan estimates that there are perhaps 1 billion planets in the universe that have harbored civilizations, but only about ten civilizations with the radio astronomy that would enable them to communicate with other civilizations. More recent calculations using low estimates for different variables suggest that we are alone in the universe ($N = 9.1 \times 10^{-11}$), while others using high estimates suggest that there could be more than 150 million advanced civilizations ($N = 1.5 \times 10^{8}$).

Either way, *observations* suggest that humanity is alone. The science fiction writer David Brin refers to this eerie situation of cosmic isolation—a kind of sensory deprivation—as the **Great Silence**.[77] Later, the economist Robin Hanson proposed an explanatory framework for the Great Silence, the central idea being that there must exist at least one **Great Filter** on the path from dead matter to advanced civilizations capable of communicating with other advanced civilizations. Hanson identifies nine major evolutionary transitions that have to obtain for a civilization to reach a communicable state:

(1) The right star system (including organics)

(2) Reproductive something (e.g. RNA)

(3) Simple (prokaryotic) single-cell life

(4) Complex (archaeatic and eukaryotic) single-cell life

(5) Sexual reproduction

(6) Multicellular life

(7) Tool-using animals with big brains

(8) Where we are now

(9) Colonization explosion.[78]

Perhaps the emergence of information-carrying, self-replicating molecules (such as ribozymes, also known as RNA enzymes) is the probability bottleneck that explains the Great Silence. After all, despite decades of research, scientists have failed to produce a single instance of abiogenesis ("life from non-life") in the laboratory, no matter how carefully they recreate the hypothesized geophysical conditions of our primordial planet. (Although some, such as Stanley Miller and Harold Urey, have managed to produce the constituents of proteins from inorganic compounds.) Or maybe the rise of intelligent tool-using animals with a high encephalization quotient (i.e., brain-to-body ratio) constitutes the Great Filter. As the biologist E.O. Wilson once suggested, "Perhaps one of the laws of evolution across inhabited planets in the universe … is that intelligence usually extinguishes itself."[79] There could also be *multiple* Great Filters between (1) and (9), with the limiting case being a Great Filter at each transition.[80]

The ultimate question for existential risk scholars is whether or not a Great Filter lies in our future. One way to evaluate this question is to look backward and consider how probable the steps before (9) are. If we find that (1) through (8) are reasonably likely, then we should conclude that a Great Filter probably lies in the future. In Hanson's words, "Optimism … regarding our future is directly pitted

against optimism regarding the ease of previous evolutionary steps. To the extent those successes were easy, our future failure to [reach technological maturity] is almost certain."[81] This is precisely why Bostrom argues that discovering single-celled organisms on Mars, if independent in origin from those on Earth, would be a crushing disappointment: it would reduce the probability of one or more Great Filters associated with (1) to (3). Similarly, finding complex organisms capable of sexual reproduction would lower the probability of Great Filters associated with (1) to (5).[82] The result would be to "shift the probability more strongly to the hypothesis that the Great Filter is ahead of us."[83] By analogy, say that conditions A, B, and C are necessary and sufficient for X to obtain. If X is failing to obtain and you know that A is almost always the case, then A probably isn't the reason for X failing, so the probability that B or C is the obstructing factor increases. Thus, we should hope to find the universe utterly vacant, since this would suggest that the Great Filter lies somewhere in our past. As Bostrom wryly declares, "Dead rocks and lifeless sands would lift my spirit."[84]

On the other hand, imagine a future in which we build supercomputers capable of simulating our evolutionary history. Imagine that such simulations begin with a "lifeless" universe but that after running a large number of them we find primitive lifeforms evolving in a majority of the universes. Depending on how high-resolution the simulations are, we could take this to infer that step (1) is not improbable. Now imagine that these single-celled creatures consistently evolve into tool-using, big-brained organisms but almost never manage to establish industrial societies. What would this imply? If scientists were to find the simulated creatures consistently evolving to a particular step between (1) and (7) but not beyond, then we would have reason for thinking that the Great Filter lies behind us. In contrast, if many of our simulations were to yield industrial societies like ours but *not* technologically mature civilizations that emit powerful signals into the heavens, colonize some portion of their

Hubble volume, or launch von Neumann probes into space, then we would have greater reason for worrying about a killer catastrophe up ahead.[85]

So, using this logic, the concept of the Great Filter can help clarify the degree to which contemporary people should be nervous about phenomena like climate change, biodiversity loss, nuclear weapons, biotechnology, synthetic biology, molecular nanotechnology, artificial intelligence, and so on. If science establishes that the evolutionary transitions behind us are relatively likely, then we should fear that doom lies in our future. (For more on the Great Filter framework and the probability of doom, see section 7.1.)

1.6 Biases and Distortions

Determining the extent to which we might be in danger requires precise and accurate thought about our evolving existential situation. Yet—at the risk of asserting a platitude—thinking clearly about the world is difficult. Our cognitive capacities are limited by the information-encoding and concept-generating mechanisms bequeathed to us by evolution and, as Bruce Tonn and Dorian Stiefel report, "most individuals' abilities to imagine the future goes 'dark' at the ten-year horizon."[86] Making matters worse, our minds are susceptible to a range of *cognitive biases* that can trick us into embracing—sometimes with great confidence—incorrect beliefs about reality. Given that the stakes are astronomically high, scholars should be especially careful to guard against the many intellectual prejudices that can distort our thinking. A short list of biases relevant to existential risk studies includes:

(i) **Conjunction fallacy**. Consider Linda, who "is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear

demonstrations."[87] Given this information, which of the following two statements is more *probable*: (a) Linda is a bank teller, or (b) Linda is a bank teller and is active in the feminist movement? When subjects are asked this question, the majority opt for (b) over (a). After all, (b) is more *representative* of Linda's description, and consequently it appears more plausible. But plausibility does not equal probability. In fact, the objectively correct answer is that (a) is more likely true than (b). Why? Because (b) contains (a), resulting in an asymmetry such that for (b) to be true, (a) must also be true, but for (a) to be true, (b) need not be true.[88]

Anytime a proposition is added to another proposition, the resulting conjunction is (as a whole) necessarily less probable than either of the two propositions individually.[89] This is because two propositions conjoined and asserted as true require more to be the case in the world (assuming the *correspondence theory of truth*).[90] Whereas (a) requires one condition to hold, (b) requires two. This loosely relates to the principle of Occam's razor, which states that when two hypothesis explain a given phenomenon, or *explanandum*, equally well, one should always choose the simpler hypothesis. As the philosopher Graham Oddie writes in the *Stanford Encyclopedia of Philosophy*, the "degree of informative content varies inversely with probability—the greater the content the less likely a theory is to be true."[91]

Now consider an alternative situation involving Linda. Which of the following is more probable: (a) Linda is active in the feminist movement, or (b) Linda is active in the feminist movement *or* is a bank teller? The correct answer now is that (b) is more probable, since it could be true *even if* Linda isn't active in the feminist movement. In other words, (b) increases the number of ways that it could be true by adding another proposition not through conjunction but disjunction. And the more that disjuncts

are added, the more probable the resulting proposition (as a whole) will be.[92]

The relevance is this: existential risk scenarios like human extinction and permanent stagnation are causally *disjunctive*. That is, they could happen as a result of asteroids *or* supervolcanoes *or* climate change *or* nuclear war *or* designer pathogens *or* superintelligence, etc. Yet the human mind "prefers" conjunctions. Consequently, we may overestimate elaborate risk scenarios while underestimating the total risk posed by a growing number of deadly threats, or we may judge elaborate arguments against certain risk scenarios to be more convincing than they are, which could leave us unnecessarily vulnerable.

(ii) **Confirmation bias**. John is a huge supporter of a politician named Zoe. Unfortunately, his close friends don't share his excitement because they believe that Zoe is a pathological liar. To convince himself that Zoe is trustworthy, John curates ten impressive instances when Zoe told a hard truth, complete with verifiable citations (e.g., from PolitiFact). Does this evidence justify his prior beliefs about Zoe's probity? No, because evaluating truth-claims requires taking into account both confirming *and* disconfirming cases—an issue we will revisit in the next section. There could, indeed, be 100 cases of Zoe offering a complete fabrication to cover up criminal acts and malfeasance, which would suggest that Zoe is duplicitous after all. The flip side of this phenomenon is the disconfirmation bias, which occurs when one spends more time scrutinizing evidence that contradicts one's preferred beliefs than evidence that supports them. For example, imagine that John's friends present the 100 instances of Zoe lying to John in an attempt to sway his opinion. Since John wants to believe that Zoe is truthful, he responds by assiduously researching every single accusation to show how each might be flawed. In contrast, he spends virtually no time ensuring that the

ten instances of Zoe stating the facts are accurate beyond a reasonable doubt.

This bias could nontrivially influence work on existential risks. For example, a stubborn optimist might spend all her time poking holes in arguments that humanity is in danger while uncritically elevating data that suggests our future is safe. The result of such tendentious research, on the optimist's part, could have catastrophic consequences if she were to persuade society to let down its guard. Alternatively, an existential risk scholar with alarmist inclinations and a career predicated on there being a high threat level might employ the exact same techniques to reach exaggerated conclusions about how risky our situation is. Both cases must be avoided, and the only way to do this is to embrace the epistemic attitude of *intellectual honesty*, which means (a) considering all the evidence, and (b) treating all the evidence the same, even when this leads to psychological disappointment.

(iii) **Observation selection effect**. This is a type of selection effect that arises from the fact that certain types of catastrophes are incompatible with the existence of observers like us. It can lead people to overestimate the probability of survival based on the empirical fact that human extinction has never before occurred. But observers like us can only ever find themselves in situations in which there are no extinction events in our species' evolutionary past. Thus, the fact that we have not yet gone extinct should not be surprising. Similarly, consider an Ultimate X-Risk that could destroy the entire universe in an instant. Can the past provide any useful information about how probable this event is? Apparently not. Whether or not an Ultimate X-Risk is extremely probable or improbable, we should expect to find ourselves in a world exactly like this one, with fully intact galaxies, stars, and planets. Both hypotheses (probable versus improbable) predict the very same observations. As Ćirković puts the point, "People often erroneously claim that we should not worry too much about

existential disasters, since none has happened in the last thousand or even million years. This fallacy needs to be dispelled."[93]

Other cognitive distortions relevant to existential risk studies include:

- **Availability bias**: This occurs when people "rely too strongly on information that is readily *available* [while ignoring] information that is less available."[94]
- **Gambler's fallacy**: "The tendency to think that future probabilities are changed by past events, when in reality they are unchanged."[95]
- **Good-story bias**: Our intuitions about the future are often shaped by popular books and movies, and thus may be biased toward exciting storylines, independent of their probability.[96]
- **Affect heuristic**: This "refers to the way in which subjective impressions of 'goodness' or 'badness' can act as a heuristic, capable of producing fast perceptual judgments, and also systematic biases."[97]
- **Motivated reasoning**: "Rather than search rationally for information that either confirms or disconfirms a particular belief, people actually seek out information that confirms what they already believe."[98]
- **Scope neglect**: This "occurs when the valuation of a problem is not valued with a multiplicative relationship to its size."[99]
- **Superiority bias**: "The belief that you are better than average in any particular metric."[100]
- **Negativity bias**: The human tendency to react more strongly to stimuli that have a negative valence.
- **Optimism bias**: The persistent belief that the future will be

better than the past and present.[101]

- **Anchoring**: "The common human tendency to rely too heavily on the first piece of information offered (the 'anchor') when making decisions."[102]
- **Base rate fallacy**: This happens when "people order information by its *perceived degree of relevance*, and let high-relevance information dominate low-relevance information."[103]
- **Hindsight bias**: "A memory distortion phenomenon by which, with the benefit of feedback about the outcome of an event, people's recalled judgments of the likelihood of that event are typically closer to the actual outcome than their original judgments were."[104]
- **Overconfidence**: This involves someone believing "that his or her judgement is better or more reliable than it objectively is."[105]

Although we won't discuss these any further here, readers are strongly encouraged to familiarize themselves more closely with these phenomena.[106]

1.7 The Epistemology of Eschatology

Eschatology: the study of the end of the world Epistemology: the theory of knowledge

Just as section 1.5 placed humanity in a larger *cosmic* context, let's now consider the broader *cultural* context in which we find ourselves. This section makes several important points that underlie and motivate the nascent field of existential risk studies, and although it may appear to delve into excessive detail, I would encourage readers not to dismiss this material too quickly.

To begin, the record of human beings claiming that their generation is the last is historically extensive—far more extensive than is generally known. The first linear **eschatological** narrative was probably invented by the ancient Persians. According to the prophet Zoroaster, also known as Zarathustra, cosmic history consists of three or four periods (depending on the tradition), each of which is exactly three millennia long. The last period culminates with the arrival of a messianic virgin-born savior, the Saoshyant, who will usher in a bodily resurrection of the dead, a Final Judgment of humanity, and an Armageddon-like war between the cosmic opposites of Good and Evil. This eschatology very likely influenced the end-times narratives of Judaism (during the Second Temple period), and consequently the two other Abrahamic religions, namely, Christianity and Islam. If this is true, which appears to be the case, then we have an argument for Zoroaster being *the most influential human to have ever existed.*[107]

Now, consider how the popular interpretation of Christian scripture known as *dispensationalism* compares to the above, albeit brief, story. According to this view, history consists of seven distinct periods called "dispensations." Contemporary humans are living in the second-to-last dispensation known as "Grace," which will conclude after Jesus briefly returns to Earth to "rapture" all the Christians, both alive and dead, who have existed since roughly 70 CE.[108] After this, a seven-year period called the Tribulation will commence, during which the Antichrist will rule a powerful governmental body like the European Union or the United Nations.[109] People will suffer immensely, especially the Jews and those who convert to Christianity after the rapture. The end of the Tribulation will be marked by the Second Coming of Christ (the *Parousia*) and the battle of Armageddon, perhaps in propinquity to the ancient town of Megiddo, Israel. Jesus will cast the Antichrist into the Lake of Fire, and the final dispensation—the Millennial Kingdom—will commence.[110] At the end of this 1,000-year period, there will be yet another great battle, this time between God and Satan,

involving the nations of Gog and Magog, followed by another bodily resurrection of the dead and one last judgment of humanity, called the "Great White Throne Judgment."[111] All true Christians will enter paradise in heaven and the unbelievers will be banished to perdition for eternity.

Paralleling this narrative in certain notable respects, some traditions in Sunni Islam prophesy that an end-of-days messianic figure called the Mahdi will appear in Mecca, Saudi Arabia, and lead an army of Muslims into an Armageddon-like battle in the small town of Dabiq in northern Syria, near Aleppo. After Armageddon, the remaining Muslim army will travel to and supernaturally conquer Constantinople (now Istanbul).[112] The Antichrist, or *Dajjal*, will then make his appearance, spreading horrible evil throughout the entire world. But his arrival, the first of Ten Major Signs of the Last Hour, will be followed by the second Major Sign, namely, the descent of Jesus on the wings of two angels. This will occur over the White Minaret of the Umayyad Mosque, in modern-day Damascus. Jesus will then chase the Antichrist to the "gate of Ludd," now called "Lod" in Israel, at which point he will kill the Antichrist. Other Major Signs will follow, most of which are quite bizarre, such as the sun rising from the West and the emergence of the ferocious killing machines Gog and Magog, whom God will utterly decimate. At the very terminus of cosmic history, God will oversee a bodily resurrection and Final Judgment of humanity. All true Muslims will enter heaven and the infidels will be cast into hell forever.[113]

There are a couple of issues worth pausing over here. First, I would argue that it is vital for existential risk scholars to understand these narratives in some detail. The reason is that they are widely believed around the planet and have shaped world history in *truly profound ways*. Consider that an incredible 41 percent of U.S. Christians in 2010 avowed that Jesus will either "definitely" or "probably" return by 2050.[114] One finds a similar prevalence of end-

times beliefs in the Muslim world, with, for example, 83 percent of Muslims in Afghanistan and 72 percent in Iraq claiming that the Mahdi will return within their lifetimes.[115]

Looking back to the origin of these faiths, both Jesus and Muhammad may have believed that the end was nigh in their own day. As the majority of New Testament scholars today maintain—following the influential theologian Albert Schweitzer—Jesus was probably a failed apocalyptic prophet who voluntarily sacrificed himself "to force the hand of God" when it became clear that the world was not about to end.[116] With respect to Islam, the historian Allen Fromherz writes that "some scholars have suggested that Islam was, from the first revelations of Muhammad, almost entirely an apocalyptic movement…. Some have even supposed that Muhammad deliberately failed to designate a successor because he predicted that the final judgment would occur after his death."[117]

Furthermore, numerous conflicts of historical significance have been greatly influenced by interpretations of Christian and Islamic eschatology—a phenomenon that I call the "clash of eschatologies."[118] For instance, as subsection 4.3.1 explores, many contemporary Islamic terrorist groups, both Sunni and Shia, are animated by "active apocalyptic" beliefs according to which they see themselves as *fervent participants in an apocalyptic narrative that is unfolding in real-time.*[119] But the plot thickens, because some of the most prominent Islamic terrorist groups today have emerged in direct response to two recent U.S.-led incursions, namely, the 1990 Gulf War and the 2003 Iraq War. And both of these may have been shaped by eschatological convictions associated with what scholars call the "Armageddon lobby" in the United States—that is, a large demographic of leaders and constituents whose political worldviews are intimately linked to dispensationalism.[120] Even more, many Islamists accuse Western forces stationed in the Middle East of being "crusaders," a term that gestures back to the religious wars of the Crusades; and as the

terrorism expert Will McCants notes, "The 100,000 European foreign fighters who flooded into Palestine under the banner of the First Crusade believed they were hastening the End of Days."[121] So, the ongoing violence in the Middle East—currently the world's epicenter of conflict—has been fueled *for centuries* by end-times beliefs held by both Christians and Muslims.

Perhaps most intriguingly, the two most consequential "secular" movements of the twentieth century, namely, Marxism and Nazism, appear to have been inspired by religious grand narratives of history. For example, Marx believed that humanity started out in a state of primitive communism (the Garden of Eden), after which we passed through stages (dispensations) like feudalism and capitalism. In the end, humanity will enter into a paradisiacal world of pure communism (heaven on Earth) thanks to the efforts of Marx (a messianic prophet), who introduced the message of communism to the proletariat. But this last step to paradise will only occur, as the historians Daniel Chirot and Clark McCauley note, after "a final, terrible revolution" (Armageddon) that will "wipe out capitalism, alienation, exploitation, and inequality" (sin).[122] Similarly, Chirot and McCauley write that

> *It was not an accident that Hitler promised a Thousand Year Reich, a millennium of perfection, similar to the thousand-year reign of goodness promised in Revelation before the return of evil, the great battle between good and evil, and the final triumph of God over Satan. The entire imagery of his Nazi Party and regime was deeply mystical, suffused with religious, often Christian, liturgical symbolism, and it appealed to a higher law, to a mission decreed by fate and entrusted to the prophet Hitler.*[123]

It is considerations like these that lead the biblical scholar and terrorism expert Frances Flannery to declare that "the Book of

Revelation has arguably been responsible for more genocide and killing in history than any other [book]." Elsewhere she claims that Revelation is

> *responsible, directly or indirectly, for massive amounts of violence. In fact, it is arguably the bloodiest book in history. Even today, groups and individuals as diverse as the Oklahoma City bombers and radical Islamist groups … have each updated the Book of Revelation to apply to their own period and causes, using it to justify violence and brutality.*[124]

Thus anxious anticipation of, and even outright elation about, the apocalypse can be found across cultural space and time.[125] This leads to a second important point: the fact that so many people have sounded the alarm bell throughout history may lead some observers to dismiss contemporary concerns from the existential risk community about global catastrophic risks. Such skeptical people might say, "Why should I believe doomsaying scientists? *Every* generation throughout history has had *somebody* claiming that their generation is the last. This is just more of the same alarmist nonsense."

But this objection is deeply misguided for reasons relating to a single crucial topic: **epistemology**. This refers to the subfield of philosophy dedicated to understanding truth, justification, and knowledge. Epistemological questions include: What constitutes truth? What conditions make a belief reasonable? Of what does knowledge consist? The most important issue for the present discussion concerns what we can call "epistemic *justification*, *warrant*, or *reasonableness*," where these terms are more or less interchangeable in this context.

The point is that *science*—our very best strategy for acquiring knowledge about the universe—is based on a highly rigorous interpretation of epistemic justification. Theories must be not merely

compatible with, but positively *supported by* some form of intersubjectively verifiable evidence.[126] And not just any evidence, but rather the *totality of evidence available at a given time*.[127] This last point is important for the following reason: imagine two competing hypotheses, A and B. Hypothesis A has, let us say, two "pieces" of evidence supporting it. Should one accept it as true, given this evidential support? The answer depends on whether hypothesis B has more than, less than, or equal to two "pieces" of evidence. If B has, for instance, 20 "pieces" of evidence in its favor, then it would be irrational to believe A. The *totality condition* of reasonable belief is a feature that many religious extremists, conspiracy theorists, and psychotic people fail to consider, thus leading them to accept unwarranted propositions that nonetheless may have some evidential support. Since humanity can't peek under the hood of reality, the best we can hope for in life is to be as reasonable as possible—that is, to construct worldviews whose interlocking beliefs are founded on objective evidence considered as a whole and constantly responsive to changes in the pool of available evidence as ongoing research uncovers new data.[128]

The further point is that, as indicated above, existential risk studies is a thoroughly *scientific* discipline. It uses the tools and methods of *rational empiricism* to map out the obstacle course of risks that civilization must navigate in the coming decades and centuries—and beyond. Even in the case of highly speculative risk scenarios such as a superintelligence takeover or a simulation shutdown, the core line of reasoning involves empirical trends, objective knowledge, and logical inferences. In contrast, the world's many religious traditions are based not on evidence but faith, and the source of knowledge comes not from observation but private revelation and testimony.[129] This makes the epistemological status of religious eschatology *fundamentally incommensurable* with that of existential risk studies, and this difference accounts for why one should listen to scientists and philosophers worried about the apocalypse but not religious folks.

To adapt a phrase from the philosopher David Hume, *the wise person always proportions her or his fears to the best available evidence, considered as a whole*. It follows that fear itself is not bad or undesirable as long as it is rational. Indeed, our best chance of surviving this century is to let what we might call *intelligent anxiety* be our guide and chaperone as we move forward. Just as long as this anxiety is motivating rather than defeatist (see section 7.2), it could be the key that unlocks our posthuman future.

Chapter 2: Our Cosmic Risk Background

2.1 Threats from Above and Below

The astrophysicist Neil deGrasse Tyson was once asked before filming a video for Big Think to briefly discuss any topic of his choosing. In deadpan fashion, Tyson intoned that "the universe is a deadly place. At every opportunity, it's trying to kill us. And so is Earth."[1] Humor aside, this gestures at a truth about our existential situation: the universe is an obstacle course of deadly hazards, and it doesn't care whether intelligent life survives or perishes. We can call this obstacle course our **cosmic risk background**. There are two general risk types within this category, namely, (a) those emerging from Earth, and (b) those hiding in the heavens. Supervolcanoes and natural pandemics are examples of the former, whereas asteroids, comets, and other astronomical phenomena are instances of the latter. Let's examine these in turn.

2.2 Supervolcanoes

To review some common geological knowledge, a volcano is an opening in Earth's surface through which magma and the dissolved gases that it contains escape—sometimes violently. Scientists have devised the volcanic explosivity index (VEI) to classify the strength of eruptions. The VEI ranges from 0 to 8, where the continuous volcanic flows on Hawaii with relatively small eruptive volumes and plume heights of less than roughly 330 feet constitute a 0 and the 1815 eruption of Mount Tambora, located on the Indonesian island of Sumbawa, constitutes a 7. (See Table A.) Let us linger on the latter for a moment. On April 5, 1815, Mount Tambora began to spew ash into

the air. Subsequent explosions were loud enough for soldiers hundreds of miles away to wonder if a war might have broken out. Five days later, a plume of smoke reached 25 miles high, propelled by three pillars of fire that eventually merged into a single column of blazing rock. Toxic ash and pumice almost eight inches wide rained down upon Sumbawa, and a tsunami crashed into the beaches of nearby islands. Dead vegetation entangled with buoyant pumice created massive "rafts" floating on the ocean, some over three miles across. An estimated 10,000 people on the island died instantly from the blast, while many more perished in the aftermath, due to starvation and disease. In fact, the word "Tambora" means "gone" in the local language.[2]

Table A. Volcanic Explosivity Index with Examples

VEI	Examples
8	Toba, 72,000 BCE; Yellowstone, 640,000 BCE "Mega-colossal" with "vast" stratospheric injections
7	Tambora, 1815 "Super-colossal" with "substantial" stratospheric injections
6	Pintabu, 1991 "Colossal" with "substantial" stratospheric injections
5	Vesuvius, 79 "Paroxysmic" with "significant" stratospheric injections
4	Calbuco, 2015 "Cataclysmic" with "definite" stratospheric injections
3	Nabro, 2011 "Catastrophic" with "possible" stratospheric injections
2	Sinabung, 2010 "Explosive" with no stratospheric injections
1	Stromboli, continuous "Gentle" with no stratospheric injections
0	Kīlauea, ongoing "Effusive" with no stratospheric injections

But the worst effects were those observed across the Northern Hemisphere a year later, during the summer of 1816. Throughout Europe, the U.S., and Asia, unusually cold weather ruined the year's crops, leading to widespread food shortages. In France, this resulted in rioting; in Ireland, where rain fell for eight weeks without a hiatus, famine and malnutrition brought about an outbreak of typhus that killed thousands; in Bengal, an epidemic of cholera emerged that, after spreading around the globe, caused tens of millions of deaths; in China, people starved and some parents even killed their children "out of mercy"; and in the United States, ice covered lakes and snow blanketed regions of the East Coast as far south as Virginia during June and July.[3] This appears to have spurred a migration of folks from

the U.S. Northeast into the American heartland, as Robert Evans notes in a *Smithsonian* article:

> *Odd as it may seem, the settling of the American heartland was apparently shaped by the eruption of a volcano 10,000 miles away. Thousands left New England for what they hoped would be a more hospitable climate west of the Ohio River. Partly as a result of such migration, Indiana became a state in 1816 and Illinois in 1818.*[4]

Perhaps most intriguingly, the anomalous weather inspired a then-unknown author named Mary Shelley, vacationing in Switzerland with the British poet Lord Byron, to write *Frankenstein*.[5] Lord Byron himself composed a poem in July of 1816 called "Darkness," which includes the lines "I had a dream, which was not all a dream. / The bright sun was extinguish'd, and the stars / Did wander darkling in the eternal space, / Rayless, and pathless, and the icy earth / Swung blind and blackening in the moonless air." This "Year Without a Summer" clearly illustrates how a large volcanic eruption can have major disruptive effects around the world.

But recall that there is one level higher on the VEI scale. This is reserved for **supervolcanic eruptions** capable of ejecting *hundreds of times* more ash into the atmosphere than Tambora did. When such an eruption occurs, sulfur dioxide is catapulted into the *stratosphere*, an atmospheric layer located above the troposphere and below the mesosphere, where the sun's light converts it into sulfuric acid. It then condenses into a layer of sulfate aerosols that reflect incoming solar radiation back into space, thereby causing Earth's skies to dim and surface temperatures to drop. The reduced photosynthesis from less sunlight can precipitate major agricultural failures lasting for years or even decades, resulting in, as the geologist Michael Rampino puts it, "widespread starvation, famine, disease, social unrest, financial collapse" and, at the extreme, "severe damage to the underpinnings of

civilization."[6] Scientists refer to this scenario as a **volcanic winter**.

Numerous supervolcanic eruptions have occurred across geological time, at least 47 of which were known to science as of 2004.[7] One of the most recent happened on the Indonesian island of Sumatra circa 73,500 BCE—in fact, volcanologists coined the term "supereruption" to describe this particular event, known as the "Toba catastrophe." It may have led to a decade of severe weather changes, with average surface temperatures in the Northern Hemisphere falling by an incredible 5.4 to 9 degrees Fahrenheit.[8] According to Rampino, up to "three-quarters of the plant species in the Northern Hemisphere perished," and other studies suggest a spike in species extinctions at the time.[9] Even more, the Toba catastrophe may have caused a severe bottleneck in the population of our ancestors, with some experts estimating as few as 500 breeding females surviving, and human population sizes shrinking to "as small as 4000 for approximately 20,000 years."[10] Thus, if the diachronic tape of anthropological history were rewound and played again, *Homo sapiens* might not have made it through the Pleistocene.

On average, supereruptions occur about once every 50,000 years. As the Geological Society of London writes, "Sooner or later a supereruption will happen on Earth and this is an issue that … demands serious attention."[11] Unfortunately, our ability to predict supervolcanic eruptions is quite poor. For example, despite "2,000 years of observations for the Italian volcano Vesuvius, and a long history of monitoring and scientific study, prediction of the timing and magnitude of the next Vesuvian eruption remains a problem."[12] Similarly, Yellowstone National Park has seen three supereruptions over the past 2 million years, each of which "produced thick ash deposits over the western and central United States."[13] Recent studies show that the magma chamber under Yellowstone is 2.5 times bigger than previously thought, making it "close to the size of the pocket when the supervolcano last erupted, 640,000 years ago." The

geoscientist James Farrell thus notes that "what we're seeing now agrees with the geologic data that we have about past eruptions. And that means there's the potential for the same type of eruption that we've seen in the past."[14] Yet we have no way of saying when this might happen.

But even if scientists *could* make accurate predictions, this might not help us *prevent* a supereruption from occurring. As the Geological Society of London observed in 2004, "Even science fiction cannot produce a credible mechanism for averting a supereruption. The point is worth repeating. No strategies can be envisaged for reducing the power of major volcanic eruptions."[15] However, this may not be entirely true today. According to the GCR expert Seth Baum, one possible strategy involves drilling "the ground around potential supervolcanoes to extract the heat, although the technological feasibility of this proposal has not yet been established."[16] He adds that "this could be a very costly project, but, if it works, it could … reduce supervolcanoes GCR."[17] Either way, the point remains that prophylactic measures are highly limited. Perhaps our best chance of survival stems from post-eruption adaptation rather than pre-eruption mitigation, an issue to which we will return in section 6.5.

Although supervolcanoes rarely become active, spewing their innards high up into the atmosphere, they warrant serious concern because of the spatiotemporal scope of their consequences. If a Tobasized supereruption were to occur tomorrow, the result could be a global or even existential catastrophe.

2.3 Natural Pandemics

Some scholars claim that the history of civilization is the history of war. While the amount of self-inflicted human suffering is truly staggering, the facts suggest that infectious diseases have thrown more people into the grave than the innumerable conflicts fought over religion, ideology, resources, and pride. Consider the fact that from

1918 to 1920 the Spanish flu outbreak killed some 50 million people, whereas "only" about 17 million people died in World War I, which lasted from 1914 to 1918. Or note that about 3 percent of the global population (in 1940) died in World War II, whereas the Plague of Justinian killed roughly 50 percent of the European population at the time (beginning in the mid-sixth century).[18] Even more striking, the Black Death of Europe and Asia may have killed a total of 200 million people, which is more than the number of deaths caused by World War II, World War I, the Mongol conquests, the Napoleonic Wars, the Vietnam War, the American Civil War, the 2003 Iraq War, and the War of 1812 *combined* (see Table B).

Table B. Number of Deaths in Various Wars

Event	# of Deaths
World War II	85,000,000
Taiping Rebellion	35,000,000
Mongol conquests	30,000,000
World War I	21,000,000
Napoleonic Wars	7,000,000
Vietnam War	3,000,000
American Civil War	1,000,000
2003 Iraq War	500,000
War of 1812	24,000
Total	182,524,000

Note: Based on higher estimates of all these conflicts

Consequently, infectious diseases like the flu, bubonic plague, and malaria have shaped world history in many important ways. For example, disease was a major factor behind the decimation of Native American populations after the arrival of Europeans, who, like most "civilized" peoples compared to their "primitive" counterparts, carried a much higher disease burden.[19] Similarly, smallpox played a role in enabling the Spanish to conquer the Aztec Empire, with it killing some 200,000 people in total and up to 75 percent of the population in

some regions.[20] The Black Death in Europe remained a public health hazard for three centuries, "with a lasting impact on the development of the economy and cultural evolution."[21] And the HIV/AIDS pandemic from 1981 to 2006 may have snuffed out up to 65 million lives around the world—not to mention the socially harmful backlash against homosexuals from religious conservatives. More than any other infectious disease, though, malaria—caused by a parasitic protozoan and spread by the flying hypodermic needles called mosquitoes—has arguably had the greatest effect on humanity. According to the World Health Organization (WHO), about half the world's population today remains vulnerable to malaria, and during 2015 alone, some 214 million people contracted this disease, resulting in ~438,000 fatalities.[22]

It is important to note that most of the deaths caused by infection throughout history have been the result of *extreme outbreaks*. As the Global Challenges Foundation writes, "Plotting historic epidemic fatalities on a log scale reveals that these tend to follow a power law with a small exponent: many plagues have been found to follow a power law with exponent 0.26." The report adds that "if this law holds for future pandemics as well, then *the majority of people who will die from epidemics will likely die from the single largest pandemic*."[23]

So, what reason do we have for expecting a pandemic to occur in the foreseeable future? Improvements in sanitation have significantly reduced the average person's exposure to pathogens, and modern medicine—in particular, vaccines and antibiotics—offer effective ways to prevent and treat infectious bugs. There are also international organizations like the WHO keeping a close and constant eye on disease outbreaks to minimize their impact, as demonstrated by the relatively successful containment of SARS and Ebola during the 2003 and 2014 epidemics, respectively. Yet these facts are counterbalanced by modern transportation systems that enable germs to travel from one continent to another at literally the speed of a jetliner, as well as

dense urban areas like slums and megacities that make it far easier for pathogens to propagate through a population. In fact, the United Nations predicts "that 66% of the global population will live in urban centers by 2050."[24] Climate change will also exacerbate the risk of pandemics, since heat waves and flooding events will bring "more opportunity for waterborne diseases such as cholera and for disease vectors such as mosquitoes in new regions." Considerations like these have led many public health experts to claim that "we are at greater risk than ever of experiencing large-scale outbreaks and global pandemics," and that "the next outbreak contender will most likely be a surprise."[25]

There are also doctor-caused, or *iatrogenic*, illnesses that could become worrisome in the future, primarily for GCR reasons.[26] As the biomedical scientist Edwin Kilbourne writes, "An unfortunate result of medical progress can be the unwitting induction of disease and disability as new treatments are tried for the first time. Therefore, it will not be surprising if the accelerated and imaginative devising of new technologies in the future proves threatening at times."[27] Consider that in the United States alone "the true number of premature deaths associated with preventable harm to patients [is] estimated at more than 400,000 per year."[28] To put this in perspective, about 595,000 Americans were projected to have died of cancer in 2016—meaning that mistakes by doctors constitute a *major cause* of death.[29] If, as Kilbourne suggests, the medical sciences advance at an accelerating (perhaps exponential) rate, iatrogenic illnesses could become even more of a problem.

Another medicine-related threat stems from **superbugs**. This refers to *multidrug-resistant bacteria*, or bacteria that can't be treated using two or more antibiotics.[30] This has global risk implications because "antibiotics are the foundation on which all modern medicine rests. Cancer chemotherapy, organ transplants, surgeries, and childbirth all rely on antibiotics to prevent infections. If you can't treat

those, then we lose the medical advances we have made in the last 50 years."[31] According to the Centers for Disease Control and Prevention (CDC), approximately 2 million people become sick as a result of superbugs each year, and some 23,000 die; but these numbers could be dwarfed by a global superbug outbreak. As the director general of the WHO Margaret Chan ominously puts it, "Antimicrobial resistance poses a fundamental threat to human health, development and security."[32]

Predicting a pandemic is extremely difficult; nonetheless, future global outbreaks are, it appears, more or less inevitable. As one commentator writes, "Experts say we are 'due' for one. When it happens, they tell us, it will probably have a greater impact on humanity than anything else currently happening in the world."[33]

2.4 Asteroids and Comets

At least one of the biggest extinction events on Earth was the result of an asteroid or comet collision. This occurred about 65 million years ago when an object ~10 kilometers across crash-landed on the Yucatan Peninsula, resulting in the extermination of all non-avian dinosaurs—an event that changed the trajectory of life by opening up new ecological niches for mammals.[34] An asteroid or comet might also have caused the devastating Permian-Triassic extinction some 251 million years ago (although some research indicates supervolcanism as the "kill mechanism"). This was the worst extinction event in planetary history, with "95 percent of all species, 53 percent of marine families, 84 percent of marine genera, and an estimated 70 percent of land species such as plants, insects and vertebrate animals" having perished.[35] There is, indeed, a startling record of large heavenly bodies wreaking mass havoc on Earth's biosphere. (See Box 3.)

As of this writing, scientists know about exactly 1,771 **potentially**

hazardous asteroids circling Earth.[36] Such objects could, by definition, obliterate a sizable region of the planet, wiping out entire cities or coastlines. For example, if an asteroid were to descend above a high-density urban center, the resulting losses could be similar to the detonation of a nuclear weapon. In the latter case, even a relatively small impact could "on the more pessimistic analyses lead to waves 4–7 [meters] high all around the [Pacific] rim, presumably with the loss of millions of lives," since "over 100 million people live within 20 m of sea level and 2 km from the ocean."[37]

Box 3. Consider a few recent close calls, beginning with the 2013 "Chelyabinsk event." This unfolded when an asteroid moving at about 42,000 miles per hour entered the atmosphere above the Russian city of Chelyabinsk, producing more light than the sun as it burned up. Numerous dashcam videos recorded the event, which damaged buildings, shattered windows, and injured nearly 1,500 people, resulting in 33 million U.S. dollars' worth of destruction. Four decades earlier, in 1972, a meteoroid "bounced" off Earth's atmosphere over the western United States, similar to the way a stone can skip across water. It came within 35 miles of Alberta, Canada, and if it had struck North America in the middle of the Cold War (note: a situation that the United States may be re-entering with Russia today) it could have initiated a retaliatory nuclear strike from the United States. This is known as the "Great Daylight Fireball." Looking back even further, an asteroid between 200 and 620 feet wide exploded over Siberia in 1908 with the energy output of a hydrogen bomb, flattening an area of forest roughly 770 square miles. Fortunately, the "Tunguska event" occurred over a region that was sparsely populated, so no one was injured.

For an impactor to destroy civilization or bring about our extinction, though, it would need to be at least 1 kilometer across.

Objects this large only strike Earth on average once every 500,000 years. If such a collision were to occur, it would kick up huge quantities of hot ash and dust into the stratosphere that would spread around the globe, blocking out incoming solar radiation. Consequently, "continental temperatures would plummet, and heat would flow from the warmer oceans onto the cooled land masses, resulting in violent, freezing winds blowing from sea to land."[38] An even higher-energy collision could bring about a global mass extinction event that would leave an indelible mark of catastrophe in fossiliferous strata. The astronomer William Napier describes this nightmare scenario as follows:

> *Regionally, the local atmosphere might simply be blown into space. A rain of perhaps 10 million boulders, metre sized and upwards, would be expected over at least continental dimensions …. Major global effects include wildfires through the incinerating effect of dust thrown around the Earth; poisoning of the atmosphere and ocean by dioxins, acid rain, sulphates and heavy metals; global warming due to water and carbon dioxide injections; followed some years later by global cooling through drastically reduced insolation, all of this happening in pitch black. The dust settling process might last a year to a decade with catastrophic effects on the land and sea food chains.*[39]

Indeed, the most commonly discussed risk associated with a large asteroid or comet impact is the possibility of an **impact winter**, similar to the volcanic winter phenomenon discussed above. This would induce global agricultural failures, mass starvation, malnutrition, and infectious disease outbreaks, all of which could cause major disruptions in the social, political, and economic foundations of civilization. At the extreme, an impact winter lasting years or decades could bring about a planetary-scale cataclysm from