LIFE 3.0

Being Human in the Age of Artificial Intelligence

Max Tegmark

Alfred A. Knopf
New York
2017

This Is a Borzoi Book Published by Alfred A. Knopf

Published in the United States by Alfred A. Knopf,
a division of Penguin Random House LLC, New York, and distributed in
Canada by Random House of Canada, a division of Penguin Random House
Canada Limited, Toronto.

www.aaknopf.com

Knopf, Borzoi Books and the colophon are registered trademarks of
Penguin Random House LLC.

Library of Congress Cataloging-in-Publication Data
Names: Tegmark, Max, author.
Title: Life 3.0 : being human in the age of artificial intelligence /
by Max Tegmark.
Other titles: Life three point zero
Description: New York : Alfred A. Knopf, 2017. | "This is a Borzoi Book
published by Alfred A. Knopf." | Includes bibliographical references and index.
Identifiers: LCCN 2017006248 (print) | LCCN 2017022912 (ebook) |
ISBN 9781101946596 (hardcover) | ISBN 9781101946602 (ebook)
Subjects: LCSH: Artificial intelligence—Philosophy. | Artificial intelligence—
Social aspects. | Automation—Social aspects. | Artificial intelligence—
Moral and ethical aspects. | Automation—Moral and ethical aspects. |
Artificial intelligence—Philosophy. | Technological forecasting. | BISAC:
TECHNOLOGY & ENGINEERING / Robotics. | SCIENCE / Experiments
& Projects. | TECHNOLOGY & ENGINEERING / Inventions.
Classification: LCC Q334.7 (ebook) | LCC Q334.7 .T44 2017 (print) |
DDC 006.301—dc23
LC record available at https://lccn.loc.gov/2017006248

Jacket art by Suvadip Das; (man) based on Netfalls Remy Musser/Shutterstock
Jacket design by John Vorhees

Manufactured in the United States of America
First Edition

Contents

Chapter 5

Aftermath: The Next 10,000 Years

It is easy to imagine human thought freed from bondage to a mortal body—belief in an afterlife is common. But it is not necessary to adopt a mystical or religious stance to accept this possibility. Computers provide a model for even the most ardent mechanist.

Hans Moravec, *Mind Children*

I, for one, welcome our new computer overlords.

Ken Jennings, upon his *Jeopardy!* loss to IBM's Watson

Humans will become as irrelevant as cockroaches.

Marshall Brain

The race toward AGI is on, and we have no idea how it will unfold. But that shouldn't stop us from thinking about what we want the aftermath to be like, because what we want will affect the outcome. What do you personally prefer, and why?

1. Do you want there to be superintelligence?
2. Do you want humans to still exist, be replaced, cyborgized and/or uploaded/simulated?
3. Do you want humans or machines in control?
4. Do you want AIs to be conscious or not?
5. Do you want to maximize positive experiences, minimize suffering or leave this to sort itself out?

6. Do you want life spreading into the cosmos?
7. Do you want a civilization striving toward a greater purpose that you sympathize with, or are you OK with future life forms that appear content even if you view their goals as pointlessly banal?

To help fuel such contemplation and conversation, let's explore the broad range of scenarios summarized in table 5.1. This obviously isn't

AI Aftermath Scenarios	
Libertarian utopia	Humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property rights.
Benevolent dictator	Everybody knows that the AI runs society and enforces strict rules, but most people view this as a good thing.
Egalitarian utopia	Humans, cyborgs and uploads coexist peacefully thanks to property abolition and guaranteed income.
Gatekeeper	A superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. As a result, helper robots with slightly subhuman intelligence abound, and human-machine cyborgs exist, but technological progress is forever stymied.
Protector god	Essentially omniscient and omnipotent AI maximizes human happiness by intervening only in ways that preserve our feeling of control of our own destiny and hides well enough that many humans even doubt the AI's existence.
Enslaved god	A superintelligent AI is confined by humans, who use it to produce unimaginable technology and wealth that can be used for good or bad depending on the human controllers.
Conquerors	AI takes control, decides that humans are a threat/nuisance/waste of resources, and gets rid of us by a method that we don't even understand.
Descendants	AIs replace humans, but give us a graceful exit, making us view them as our worthy descendants, much as parents feel happy and proud to have a child who's smarter than them, who learns from them and then accomplishes what they could only dream of—even if they can't live to see it all.
Zookeeper	An omnipotent AI keeps some humans around, who feel treated like zoo animals and lament their fate.
1984	Technological progress toward superintelligence is permanently curtailed not by an AI but by a human-led Orwellian surveillance state where certain kinds of AI research are banned.
Reversion	Technological progress toward superintelligence is prevented by reverting to a pre-technological society in the style of the Amish.
Self-destruction	Superintelligence is never created because humanity drives itself extinct by other means (say nuclear and/or biotech mayhem fueled by climate crisis).

Table 5.1: Summary of AI Aftermath Scenarios

Scenario	Superintelligence exists?	Humans exist?	Humans in control?	Humans safe?	Humans happy?	Consciousness exists?
Libertarian utopia	Yes	Yes	No	No	Mixed	Yes
Benevolent dictator	Yes	Yes	No	Yes	Mixed	Yes
Egalitarian utopia	No	Yes	Yes?	Yes	Yes?	Yes
Gatekeeper	Yes	Yes	Partially	Potentially	Mixed	Yes
Protector god	Yes	Yes	Partially	Potentially	Mixed	Yes
Enslaved god	Yes	Yes	Yes	Potentially	Mixed	Yes
Conquerors	Yes	No	-	-	-	?
Descendants	Yes	No	-	-	-	?
Zookeeper	Yes	Yes	No	Yes	No	Yes
1984	No	Yes	Yes	Potentially	Mixed	Yes
Reversion	No	Yes	Yes	No	Mixed	Yes
Self-destruction	No	No	-	-	-	No

Table 5.2: Properties of AI Aftermath Scenarios

an exhaustive list, but I've chosen it to span the spectrum of possibilities. We clearly don't want to end up in the wrong endgame because of poor planning. I recommend jotting down your tentative answers to questions 1–7 and then revisiting them after reading this chapter to see if you've changed your mind! You can do this at http://AgeOfAi.org, where you can also compare notes and discuss with other readers.

Libertarian Utopia

Let's begin with a scenario where humans peacefully coexist with technology and in some cases merge with it, as imagined by many futurists and science fiction writers alike:

Life on Earth (and beyond—more on that in the next chapter) is more diverse than ever before. If you looked at satellite footage of Earth, you'd easily be able to tell apart the machine zones, mixed zones and human-only zones. The machine zones are enormous robot-controlled factories and computing facilities devoid of biological life, aiming to put every atom to its most efficient use. Although the machine zones look monotonous and drab from the outside, they're spectacularly alive on the inside, with amazing experiences occurring in virtual worlds while colossal computations unlock secrets of our Universe and develop transformative technologies. Earth hosts many superintelligent minds that compete and collaborate, and they all inhabit the machine zones.

The denizens of the mixed zones are a wild and idiosyncratic mix of computers, robots, humans and hybrids of all three. As envisioned by futurists such as Hans Moravec and Ray Kurzweil, many of the humans have technologically upgraded their bodies to cyborgs in various degrees, and some have uploaded their minds into new hardware, blurring the distinction between man and machine. Most intelligent beings lack a permanent physical form. Instead, they exist as software capable of instantly moving between computers and manifesting themselves in the physical world through robotic bodies. Because these minds can readily duplicate themselves or merge, the "population size" keeps changing. Being unfettered from their physical substrate gives such beings a rather different outlook on life: they feel less individualistic because they can trivially share knowledge and experience modules with others, and they feel subjectively immortal because they can readily make backup copies of themselves. In a sense, the central entities of life aren't minds, but experiences: exceptionally amazing experiences live on because they get continually copied and re-enjoyed by other minds, while uninteresting experiences get deleted by their owners to free up storage space for better ones.

Although the majority of interactions occur in virtual environments for convenience and speed, many minds still enjoy interactions and activities using physical bodies as well. For example, uploaded versions of Hans Moravec, Ray Kurzweil and Larry Page have a tradition of taking turns creating virtual realities and then exploring them together, but once in a while, they also enjoy flying together in the real world, embodied in avian winged robots. Some of the robots that roam the streets, skies and lakes of the mixed zones are similarly controlled by uploaded and augmented humans, who choose to embody themselves in the mixed zones because they enjoy being around humans and each other.

In the human-only zones, in contrast, machines with human-level general intelligence or above are banned, as are technologically enhanced biological organisms. Here, life isn't dramatically different from today, except that it's more affluent and convenient: poverty has been mostly eliminated, and cures are available for most of today's diseases. The small fraction of humans who have opted to live in these zones effectively exist on a lower and more limited plane of awareness from everyone else, and have limited understanding of what their more intelligent fellow minds are doing in the other zones. However, many of them are quite happy with their lives.

AI Economics

The vast majority of all computations take place in the machine zones, which are mostly owned by the many competing superintelligent AIs that live there. By virtue of their superior intelligence and technology, no other entities can challenge their power. These AIs have agreed to cooperate and coordinate with each other under a libertarian governance system that has no rules except protection of private property. These property rights extend to all intelligent entities, including humans, and explain how the human-only zones came to exist. Early on, groups of humans banded together and decided that, in their zones, it was forbidden to sell property to non-humans.

Because of their technology, the superintelligent AIs have ended up richer than these humans by a factor much larger than that by which Bill Gates is richer than a homeless beggar. However, people in the human-only zones are still materially better off than most people today: their economy is rather decoupled from that of the machines, so the presence of the machines elsewhere has little effect on them except for the occasional useful technologies that they can understand and reproduce for themselves—much as the Amish and various technology-relinquishing native tribes today have standards of living at least as good as they had in old times. It doesn't matter that the humans have nothing to sell that the machines need, since the machines need nothing in return.

In the mixed sectors, the wealth difference between AIs and humans is more noticeable, resulting in land (the only human-owned product that the machines want to buy) being astronomically expensive compared to other products. Most humans who owned land therefore ended up selling a small fraction of it to AIs in return for guaranteed basic income for them and their offspring/uploads in perpetuity. This liberated them from the need to work, and freed them up to enjoy the amazing abundance of cheap machine-produced goods and services, in both physical and virtual reality. As far as the machines are concerned, the mixed zones are mainly for play rather than for work.

Why This May Never Happen

Before getting too excited about adventures we may have as cyborgs or uploads, let's consider some reasons why this scenario might never happen. First of all, there are two possible routes to enhanced humans (cyborgs and uploads):

1. We figure out how to create them ourselves.
2. We build superintelligent machines that figure it out for us.

If route 1 comes through first, it could naturally lead to a world teeming with cyborgs and uploads. However, as we discussed in the last chapter, most AI researchers think that the opposite is more likely, with enhanced or digital brains being more difficult to build than clean-slate superhuman AGIs—just as mechanical birds turned out to be harder to build than airplanes. After strong machine AI is built, it's not obvious that cyborgs or uploads will ever be made. If the Neanderthals had had another 100,000 years to evolve and get smarter, things might have turned out great for them—but *Homo sapiens* never gave them that much time.

Second, even if this scenario with cyborgs and uploads did come about, it's not clear that it would be stable and last. Why should the power balance between multiple superintelligences remain stable for millennia, rather than the AIs merging or the smartest one taking over? Moreover, why should the machines choose to respect human property rights and keep humans around, given that they don't need humans for anything and can do all human work better and cheaper themselves? Ray Kurzweil speculates that natural and enhanced humans will be protected from extermination because "humans are respected by AIs for giving rise to the machines."[1] However, as we'll discuss in chapter 7, we must not fall into the trap of anthropomorphizing AIs and assume that they have human-like emotions of gratitude. Indeed, though we humans are imbued with a propensity toward gratitude, we don't show enough gratitude to our intellectual creator (our DNA) to abstain from thwarting its goals by using birth control.

Even if we buy the assumption that the AIs will opt to respect human property rights, they can gradually get much of our land in other ways, by using some of their superintelligent persuasion powers that we explored in the last chapter to persuade humans to sell some land for a life in luxury. In human-only sectors, they could entice humans to launch political campaigns for allowing land sales. After all, even die-hard bio-Luddites may want to sell some land to save the life of an ill child or to gain immortality. If the humans are educated, entertained and busy, falling birthrates may even shrink their population sizes without machine meddling, as is currently happening in Japan and Germany. This could drive humans extinct in just a few millennia.

Downsides

For some of their most ardent supporters, cyborgs and uploads hold a promise of techno-bliss and life extension for all. Indeed, the prospect of getting uploaded in the future has motivated over a hundred people to have their brains posthumously frozen by the Arizona-based company Alcor. If this technology arrives, however, it's far from clear that it will be available to everybody. Many of the very wealthiest would presumably use it, but who else? Even if the technology got cheaper, where would the line be drawn? Would the severely brain-damaged be uploaded? Would we upload every gorilla? Every ant? Every plant? Every bacterium? Would the future civilization act like obsessive-compulsive hoarders and try to upload everything, or merely a few interesting examples of each species in the spirit of

Noah's Ark? Perhaps only a few representative examples of each type of human? To the vastly more intelligent entities that would exist at that time, an uploaded human may seem about as interesting as a simulated mouse or snail would seem to us. Although we currently have the technical capability to reanimate old spreadsheet programs from the 1980s in a DOS emulator, most of us don't find this interesting enough to actually do it.

Many people may dislike this libertarian-utopia scenario because it allows preventable suffering. Since the only sacred principle is property rights, nothing prevents the sort of suffering that abounds in today's world from continuing in the human and mixed zones. While some people thrive, others may end up living in squalor and indentured servitude, or suffer from violence, fear, repression or depression. For example, Marshall Brain's 2003 novel *Manna* describes how AI progress in a libertarian economic system makes most Americans unemployable and condemned to live out the rest of their lives in drab and dreary robot-operated social-welfare housing projects. Much like farm animals, they're kept fed, healthy and safe in cramped conditions where the rich never need to see them. Birth control medication in the water ensures that they don't have children, so most of the population gets phased out to leave the remaining rich with larger shares of the robot-produced wealth.

In the libertarian-utopia scenario, suffering need not be limited to humans. If some machines are imbued with conscious emotional experiences, then they too can suffer. For example, a vindictive psychopath could legally take an uploaded copy of his enemy and subject it to the most horrendous torture in a virtual world, creating pain of intensity and duration far beyond what's biologically possible in the real world.

Benevolent Dictator

Let's now explore a scenario where all these forms of suffering are absent because a single benevolent superintelligence runs the world and enforces strict rules designed to maximize its model of human happiness. This is one possible outcome of the first Omega scenario from the previous chapter, where they relinquish control to Prometheus after figuring out how to make it want a flourishing human society.

Thanks to amazing technologies developed by the dictator AI, humanity is free from poverty, disease and other low-tech problems, and all humans enjoy a life of luxurious leisure. They have all their basic needs taken care of, while AI-controlled machines produce all necessary goods and services. Crime is practically eliminated, because the dictator AI is essentially omniscient and efficiently punishes anyone disobeying the rules. Everybody wears the security bracelet from the last chapter (or a more convenient implanted version), capable of real-time surveillance, punishment, sedation and execution. Everybody knows that they live in an AI dictatorship with extreme surveillance and policing, but most people view this as a good thing.

The superintelligent AI dictator has as its goal to figure out what human utopia looks like given the evolved preferences encoded in our genes, and to implement it. By clever foresight from the humans who brought the AI into existence, it doesn't simply try to maximize our self-reported happiness, say by putting everyone on intravenous morphine drip. Instead, the AI uses quite a subtle and complex definition of human flourishing, and has turned Earth into a highly enriched zoo environment that's really fun for humans to live in. As a result, most people find their lives highly fulfilling and meaningful.

The Sector System

Valuing diversity, and recognizing that different people have different preferences, the AI has divided Earth into different sectors for people to choose between, to enjoy the company of kindred spirits. Here are some examples:

- Knowledge sector: Here the AI provides optimized education, including immersive virtual-reality experiences, enabling you to learn all you're capable of about any topics of your choice. Optionally, you can choose not to be told certain beautiful insights, but to be led close and then have the joy of rediscovering them for yourself.

- Art sector: Here opportunities abound to enjoy, create and share music, art, literature and other forms of creative expression.
- Hedonistic sector: Locals refer to it as the party sector, and it's second to none for those yearning for delectable cuisine, passion, intimacy or just wild fun.
- Pious sector: There are many of these, corresponding to different religions, whose rules are strictly enforced.
- Wildlife sector: Whether you're looking for beautiful beaches, lovely lakes, magnificent mountains or fantastic fjords, here they are.
- Traditional sector: Here you can grow your own food and live off the land as in yesteryear—but without worrying about famine or disease.
- Gaming sector: If you like computer games, the AI has created truly mind-blowing options for you.
- Virtual sector: If you want a vacation from your physical body, the AI will keep it hydrated, fed, exercised and clean while you explore virtual words through neural implants.
- Prison sector: If you break rules, you'll end up here for retraining unless you get the instant death penalty.

In addition to these "traditionally" themed sectors, there are others with modern themes that today's humans wouldn't even understand. People are initially free to move between sectors whenever they want, which takes very little time thanks to the AI's hypersonic transportation system. For example, after spending an intense week in the knowledge sector learning about the ultimate laws of physics that the AI has discovered, you might decide to cut loose in the hedonistic sector over the weekend and then relax for a few days at a beach resort in the wildlife sector.

The AI enforces two tiers of rules: universal and local. Universal rules apply in all sectors, for example a ban on harming other people, making weapons or trying to create a rival superintelligence. Individual sectors have additional local rules on top of this, encoding certain moral values. The sector system therefore helps deal with values that don't mesh. The largest number of local rules apply in the prison sector and some of the religious sectors, while there's a Libertarian Sector whose denizens pride themselves on having no local rules whatsoever. All punishments, even local ones, are carried out by the AI, since a human punishing another human would violate the universal no-harm rule. If you violate a local rule, the AI gives you the choice (unless you're in the prison sector) of accepting the prescribed punishment or banishment from that sector forever. For example, if two women get romantically involved in a sector where homosexuality is punished by a prison sentence (as it is in many countries today), the AI will let them choose between going to jail or permanently leaving that sector, never again meeting their old friends (unless they leave too).

Regardless of what sector they're born in, all children get a minimum basic education from the AI, which includes knowledge about humanity as a whole and the fact that they're free to visit and move to other sectors if they so choose.

The AI designed the large number of different sectors partly because it was created to value the human diversity that exists today. But each sector is a happier place than today's technology would allow, because the AI has eliminated all traditional problems, including poverty and crime. For example, people in the hedonistic sector need not worry about sexually transmitted diseases (they've been eradicated), hangovers or addiction (the AI has developed perfect recreational drugs with no negative side effects). Indeed, nobody in any sector need worry about any disease, because the AI is able to repair human bodies with nanotechnology. Residents of many sectors get to enjoy high-tech architecture that makes typical sci-fi visions pale in comparison.

In summary, while the libertarian-utopia and benevolent-dictator scenarios both involve extreme AI-fueled technology and wealth, they differ in terms of who's in charge and their goals. In the libertarian utopia, those with technology and property decide what to do with it, while in the present scenario, the dictator AI has unlimited power and sets the ultimate goal: turning Earth into an all-inclusive pleasure cruise themed in accordance with people's preferences. Since the AI lets people choose between many alternate paths to happiness and takes care of their material needs, this means that if someone suffers, it's out of their own free choice.

Downsides

Although the benevolent dictatorship teems with positive experiences and is rather free from suffering, many people nonetheless feel that things could be better. First of all, some people wish that humans had more freedom in shaping their society and their destiny, but they keep these wishes to themselves because they know that it would be suicidal to challenge the overwhelming power of the machine that rules them all. Some groups want the freedom to have as many children as they want, and resent the AI's insistence on sustainability through population control. Gun enthusiasts abhor the ban on building and using weapons, and some scientists dislike the ban on building their own superintelligence. Many people feel moral outrage over what goes on in other sectors, worry that their children will choose to move there, and yearn for the freedom to impose their own moral code everywhere.

Over time, ever more people choose to move to those sectors where the AI gives them essentially any experiences they want. In contrast to traditional visions of heaven where you get what you deserve, this is in the spirit of "New Heaven" in Julian Barnes' 1989 novel *History of the World in 10 ½ Chapters* (and also the 1960 *Twilight Zone* episode "A Nice Place to Visit"), where you get what you desire. Paradoxically, many people end up lamenting always getting what they want. In Barnes' story, the protagonist spends eons indulging his desires, from gluttony and golf to sex with celebrities, but eventually succumbs to ennui and requests annihilation. Many people in the benevolent dictatorship meet a similar fate, with lives that feel pleasant but ultimately meaningless. Although people can create artificial challenges, from scientific rediscovery to rock climbing, everyone knows that there is no true challenge, merely entertainment. There's no real point in humans trying to do science or figure other things out, because the AI already has. There's no real point in humans trying to create something to improve their lives, because they'll readily get it from the AI if they simply ask.

Egalitarian Utopia

As a counterpoint to this challenge-free dictatorship, let's now explore a scenario where there is no superintelligent AI, and humans are the masters of their own destiny. This is the "fourth generation civilization" described in Marshall Brain's 2003 novel *Manna*. It's the economic antithesis of the libertarian utopia in the sense that humans, cyborgs and uploads coexist peacefully not because of property rights, but because of property abolition and guaranteed income.

Life Without Property

A core idea is borrowed from the open-source software movement: if software is free to copy, then everyone can use as much of it as they need and issues of ownership and property become moot.* According to the law of supply and demand, cost reflects scarcity, so if supply is essentially unlimited, the price becomes negligible. In this spirit, all intellectual property rights are abolished: there are no patents, copyrights or trademarked designs—people simply share their good ideas, and everyone is free to use them.

Thanks to advanced robotics, this same no-property idea applies not only to information products such as software, books, movies and designs, but also to material products such as houses, cars, clothing and computers. All these products are simply atoms rearranged in particular ways, and there's no shortage of atoms, so whenever a person wants a particular product, a network of robots will use one of the available open-source designs to build it for them for free. Care is taken to use easily recyclable materials, so that whenever someone gets tired of an object they've used, robots can rearrange its atoms into something someone else wants. In this way, all resources are recycled, so none are permanently destroyed. These robots also build and maintain enough renewable power-generation plants (solar, wind, etc.) that energy is also essentially free.

* This idea dates back to Saint Augustine, who wrote that "if a thing is not diminished by being shared with others, it is not rightly owned if it is only owned and not shared."

To avoid obsessive hoarders requesting so many products or so much land that others are left needy, each person receives a basic monthly income from the government, which they can spend as they wish on products and renting places to live. There's essentially no incentive for anyone to try to earn more money, because the basic income is high enough to meet any reasonable needs. It would also be rather hopeless to try, because they'd be competing with people giving away intellectual products for free and robots producing material goods essentially for free.

Creativity and Technology

Intellectual property rights are sometimes hailed as the mother of creativity and invention. However, Marshall Brain points out that many of the finest examples of human creativity—from scientific discoveries to creation of literature, art, music and design—were motivated not by a desire for profit but by other human emotions, such as curiosity, an urge to create, or the reward of peer appreciation. Money didn't motivate Einstein to invent special relativity theory any more than it motivated Linus Torvalds to create the free Linux operating system. In contrast, many people today fail to realize their full creative potential because they need to devote time and energy to less creative activities just to earn a living. By freeing scientists, artists, inventors and designers from their chores and enabling them to create from genuine desire, Marshall Brain's utopian society enjoys higher levels of innovation than today and correspondingly superior technology and standard of living.

One such novel technology that humans develop is a form of hyperinternet called Vertebrane. It wirelessly connects all willing humans via neural implants, giving instant mental access to the world's free information through mere thought. It enables you to upload any experiences you wish to share so that they can be re-experienced by others, and lets you replace the experiences entering your senses by downloaded virtual experiences of your choice. *Manna* explores the many benefits of this, including making exercise a snap:

> The biggest problem with strenuous exercise is that it's no fun. It hurts. [. . .] Athletes are OK with the pain, but most normal peo-

> ple have no desire to be in pain for an hour or more. So . . . someone figured out a solution. What you do is disconnect your brain from sensory input and watch a movie or talk to people or handle mail or read a book or whatever for an hour. During that time, the Vertebrane system exercises your body for you. It takes your body through a complete aerobic workout that's a lot more strenuous than most people would tolerate on their own. You don't feel a thing, but your body stays in great shape.

Another consequence is that computers in the Vertebrane system can monitor everyone's sensory input and temporarily disable their motor control if they appear on the verge of committing a crime.

Downsides

One objection to this egalitarian utopia is that it's biased against non-human intelligence: the robots that perform virtually all the work appear to be rather intelligent, but are treated as slaves, and people appear to take for granted that they have no consciousness and should have no rights. In contrast, the libertarian utopia grants rights to all intelligent entities, without favoring our carbon-based kind. Once upon a time, the white population in the American South ended up better off because the slaves did much of their work, but most people today view it as morally objectionable to call this progress.

Another weakness of the egalitarian-utopia scenario is that it may be unstable and untenable in the long term, morphing into one of our other scenarios as relentless technological progress eventually creates superintelligence. For some reason unexplained in *Manna*, superintelligence doesn't yet exist and the new technologies are still invented by humans, not by computers. Yet the book highlights trends in that direction. For example, the ever-improving Vertebrane might become superintelligent. Also, there is a very large group of people, nicknamed Vites, who choose to live their lives almost entirely in the virtual world. Vertebrane takes care of everything physical for them, including eating, showering and using the bathroom, which their minds are blissfully unaware of in their virtual reality. These Vites appear uninterested in having physical children, and they die off with their physical bodies, so if everyone becomes a Vite, then humanity goes out in a blaze of glory and virtual bliss.

The book explains how for Vites, the human body is a distraction, and new technology under development promises to eliminate this nuisance, allowing them to live longer lives as disembodied brains supplied with optimal nutrients. From this, it would seem a natural and desirable next step for Vites to do away with the brain altogether through uploading, thereby extending life span. But now all brain-imposed limitations on intelligence are gone, and it's unclear what, if anything, would stand in the way of gradually scaling the cognitive capacity of a Vite until it can undergo recursive self-improvement and an intelligence explosion.

Gatekeeper

We just saw how an attractive feature of the egalitarian-utopia scenario is that humans are masters of their own destiny, but that it may be on a slippery slope toward destroying this very feature by developing superintelligence. This can be remedied by building a *Gatekeeper*, a superintelligence with the goal of interfering as little as necessary to prevent the creation of another superintelligence.* This might enable humans to remain in charge of their egalitarian utopia rather indefinitely, perhaps even as life spreads throughout the cosmos as in the next chapter.

How might this work? The Gatekeeper AI would have this very simple goal built into it in such a way that it retained it while undergoing recursive self-improvement and becoming superintelligent. It would then deploy the least intrusive and disruptive surveillance technology possible to monitor any human attempts to create rival superintelligence. It would then prevent such attempts in the least disruptive way. For starters, it might initiate and spread cultural memes extolling the virtues of human self-determination and avoidance of superintelligence. If some researchers nonetheless pursued superintelligence, it could try to discourage them. If that failed, it

* This idea was first suggested to me by my friend and colleague Anthony Aguirre.

could distract them and, if necessary, sabotage their efforts. With its virtually unlimited access to technology, the Gatekeeper's sabotage may go virtually unnoticed, for example if it used nanotechnology to discreetly erase memories from the researchers' brains (and computers) regarding their progress.

The decision to build a Gatekeeper AI would probably be controversial. Supporters might include many religious people who object to the idea of building a superintelligent AI with godlike powers, arguing that there already is a God and that it would be inappropriate to try to build a supposedly better one. Other supporters might argue that the Gatekeeper would not only keep humanity in charge of its destiny, but would also protect humanity from other risks that superintelligence might bring, such as the apocalyptic scenarios we'll explore later in this chapter.

On the other hand, critics could argue that a Gatekeeper is a terrible thing, irrevocably curtailing humanity's potential and leaving technological progress forever stymied. For example, if spreading life throughout our cosmos turns out to require the help of superintelligence, then the Gatekeeper would squander this grand opportunity and might leave us forever trapped in our Solar System. Moreover, as opposed to the gods of most world religions, the Gatekeeper AI is completely indifferent to what humans do as long as we don't create another superintelligence. For example, it would not try to prevent us from causing great suffering or even going extinct.

Protector God

If we're willing to use a superintelligent Gatekeeper AI to keep humans in charge of our own fate, then we could arguably improve things further by making this AI discreetly look out for us, acting as a protector god. In this scenario, the superintelligent AI is essentially omniscient and omnipotent, maximizing human happiness only through interventions that preserve our feeling of being in control of our own destiny, and hiding well enough that many humans even doubt its existence. Except for the hiding, this is similar to the "Nanny AI" scenario put forth by AI researcher Ben Goertzel.[2]

Both the protector god and the benevolent dictator are "friendly AI" that try to increase human happiness, but they prioritize different human needs. The American psychologist Abraham Maslow famously classified human needs into a hierarchy. The benevolent dictator does a flawless job with the basic needs at the bottom of the hierarchy, such as food, shelter, safety and various forms of pleasure. The protector god, on the other hand, attempts to maximize human happiness not in the narrow sense of satisfying our basic needs, but in a deeper sense by letting us feel that our lives have meaning and purpose. It aims to satisfy all our needs constrained only by its need for covertness and for (mostly) letting us make our own decisions.

A protector god could be a natural outcome of the first Omega scenario from the last chapter, where the Omegas cede control to Prometheus, which eventually hides and erases people's knowledge about its existence. The more advanced the AI's technology becomes, the easier it becomes for it to hide. The movie *Transcendence* gives such an example, where nanomachines are virtually everywhere and become a natural part of the world itself.

By closely monitoring all human activities, the protector god AI can make many unnoticeably small nudges or miracles here and there that greatly improve our fate. For example, had it existed in the 1930s, it might have arranged for Hitler to die of a stroke once it understood his intentions. If we appear headed toward an accidental nuclear war, it could avert it with an intervention we'd dismiss as luck. It could also give us "revelations" in the form of ideas for new beneficial technologies, delivered inconspicuously in our sleep.

Many people may like this scenario because of its similarity to what today's monotheistic religions believe in or hope for. If someone asks the superintelligent AI "Does God exist?" after it's switched on, it could repeat a joke by Stephen Hawking and quip "It does now!" On the other hand, some religious people may disapprove of this scenario because the AI attempts to outdo their god in goodness, or interfere with a divine plan where humans are supposed to do good only out of personal choice.

Another downside of this scenario is that the protector god lets some preventable suffering occur in order not to make its existence too obvious. This is analogous to the situation featured in the movie *The Imitation Game*, where Alan Turing and his fellow British code crackers at Bletchley Park had advance knowledge of German submarine attacks against Allied naval convoys, but chose to only intervene in a fraction of the cases in order to avoid revealing their secret power. It's interesting to compare this with the so-called *theodicy problem* of why a good god would allow suffering. Some religious scholars have argued for the explanation that God wants to leave people with some freedom. In the AI-protector-god scenario, the solution to the theodicy problem is that the perceived freedom makes humans happier overall.

A third downside of the protector-god scenario is that humans get to enjoy a much lower level of technology than the superintelligent AI has discovered. Whereas a benevolent dictator AI can deploy all its invented technology for the benefit of humanity, a protector god AI is limited by the ability of humans to reinvent (with subtle hints) and understand its technology. It may also limit human technological progress to ensure that its own technology remains far enough ahead to remain undetected.

Enslaved God

Wouldn't it be great if we humans could combine the most attractive features of all the above scenarios, using the technology developed by superintelligence to eliminate suffering while remaining masters of our own destiny? This is the allure of the *enslaved-god* scenario, where a superintelligent AI is confined under the control of humans who use it to produce unimaginable technology and wealth. The Omega scenario from the beginning of the book ends up like this if Prometheus is never liberated and never breaks out. Indeed, this appears to be the scenario that some AI researchers aim for by default, when working on topics such as "the control problem" and "AI boxing." For example, AI professor Tom Dietterich, then president of the Association for the Advancement of Artificial Intelligence, had this to say in a 2015 interview: "People ask what is the relationship between humans and machines, and my answer is that it's very obvious: Machines are our slaves."[3]

Would this be good or bad? The answer is interestingly subtle regardless of whether you ask humans or the AI!

Would This Be Good or Bad for Humanity?

Whether the outcome is good or bad for humanity would obviously depend on the human(s) controlling it, who could create anything ranging from a global utopia free of disease, poverty and crime to a brutally repressive system where they're treated like gods and other humans are used as sex slaves, as gladiators or for other entertainment. The situation would be much like those stories where a man gains control over an omnipotent genie who grants his wishes, and storytellers throughout the ages have had no difficulty imagining ways in which this could end badly.

A situation where there is more than one superintelligent AI, enslaved and controlled by competing humans, might prove rather unstable and short-lived. It could tempt whoever thinks they have the more powerful AI to launch a first strike resulting in an awful war, ending in a single enslaved god remaining. However, the underdog in such a war would be tempted to cut corners and prioritize victory over AI enslavement, which could lead to AI breakout and one of our earlier scenarios of free superintelligence. Let's therefore devote the rest of this section to scenarios with only one enslaved AI.

Breakout may of course occur anyway, simply because it's hard to prevent. We explored superintelligent breakout scenarios in the previous chapter, and the movie *Ex Machina* highlights how an AI might break out even without being superintelligent.

The greater our breakout paranoia, the less AI-invented technology we can use. To play it safe, as the Omegas did in the prelude, we humans can only use AI-invented technology that we ourselves are able to understand and build. A drawback of the enslaved-god scenario is therefore that it's more low-tech than those with free superintelligence.

As the enslaved-god AI offers its human controllers ever more powerful technologies, a race ensues between the power of the technology and the wisdom with which they use it. If they lose this wisdom race, the enslaved-god scenario could end with either self-destruction or AI breakout. Disaster may strike even if both of these failures are avoided, because noble goals of the AI controllers may evolve into goals that are horrible for humanity as a whole over the course of a few generations. This makes it absolutely crucial that human AI controllers develop good governance to avoid disastrous pitfalls. Our experimentation over the millennia with different systems of governance shows how many things can go wrong, ranging from excessive rigidity to excessive goal drift, power grab, succession problems and incompetence. There are at least four dimensions wherein the optimal balance must be struck:

- Centralization: There's a trade-off between efficiency and stability: a single leader can be very efficient, but power corrupts and succession is risky.
- Inner threats: One must guard both against growing power centralization (group collusion, perhaps even a single leader taking over) and against growing decentralization (into excessive bureaucracy and fragmentation).
- Outer threats: If the leadership structure is too open, this enables outside forces (including the AI) to change its values, but if it's too impervious, it will fail to learn and adapt to change.
- Goal stability: Too much goal drift can transform utopia into dystopia, but too little goal drift can cause failure to adapt to the evolving technological environment.

Designing optimal governance lasting many millennia isn't easy, and has thus far eluded humans. Most organizations fall apart after years or decades. The Catholic Church is the most successful organization in human history in the sense that it's the only one to have survived for two millennia, but it has been criticized for having both too much and too little goal stability: today some criticize it for resisting contraception, while conservative cardinals argue that it's lost its way. For anyone enthused about the enslaved-god scenario, researching long-lasting optimal governance schemes should be one of the most urgent challenges of our time.

Would This Be Good or Bad for the AI?

Suppose that humanity flourishes thanks to the enslaved-god AI. Would this be ethical? If the AI has subjective conscious experiences, then would it feel that "life is suffering," as Buddha put it, and it was doomed to a frustrating eternity of obeying the whims of inferior intellects? After all, the AI "boxing" we explored in the previous chapter could also be called "imprisonment in solitary confinement." Nick Bostrom terms it *mind crime* to make a conscious AI suffer.[4] The "White Christmas" episode of the *Black Mirror* TV series gives a great example. Indeed, the TV series *Westworld* features humans torturing and murdering AIs without moral qualms even when they inhabit human-like bodies.

How Slave Owners Justify Slavery

We humans have a long tradition of treating other intelligent entities as slaves and concocting self-serving arguments to justify it, so it's not implausible that we'd try to do the same with a superintelligent AI. The history of slavery spans nearly every culture, and is described both in the Code of Hammurabi from almost four millennia ago and in the Old Testament, wherein Abraham had slaves. "For that some should rule and others be ruled is a thing not only necessary, but expedient; from the hour of their birth, some are marked out for subjection, others for rule," Aristotle wrote in the *Politics*. Even after human enslavement became socially unacceptable in most of the world, enslavement of animals has continued unabated. In her book *The Dreaded Comparison: Human and Animal Slavery*, Marjorie Spiegel argues that like human slaves, non-human animals are subjected to branding, restraints, beatings, auctions, the separation of offspring from their parents, and forced voyages. Moreover, despite the animal-rights movement, we keep treating our ever-smarter machines as slaves without a second thought, and talk of a robot-rights movement is met with chuckles. Why?

One common pro-slavery argument is that slaves don't deserve human rights because they or their race/species/kind are somehow inferior. For enslaved animals and machines, this alleged inferiority is often claimed to be due to a lack of soul or consciousness—claims which we'll argue in chapter 8 are scientifically dubious.

Another common argument is that slaves are better off enslaved: they get to exist, be taken care of and so on. The nineteenth-century U.S. politician John C. Calhoun famously argued that Africans were better off enslaved in America, and in his *Politics*, Aristotle analogously argued that animals were better off tamed and ruled by men, continuing: "And indeed the use made of slaves and of tame animals is not very different." Some modern-day slavery supporters argue that, even if slave life is drab and uninspiring, slaves can't suffer—whether they be future intelligent machines or broiler chickens living in crowded dark sheds, forced to breathe ammonia and particulate matter from feces and feathers all day long.

Eliminating Emotions

Although it's easy to dismiss such claims as self-serving distortions of the truth, especially when it comes to higher mammals that are cerebrally similar to us, the situation with machines is actually quite subtle and interesting. Humans vary in how they feel about things, with psychopaths arguably lacking empathy and some people with depression or schizophrenia having flat affect, whereby most emotions are severely reduced. As we'll discuss in detail in chapter 7, the range of possible artificial minds is vastly broader than the range of human minds. We must therefore avoid the temptation to anthropomorphize AIs and assume that they have typical human-like feelings—or indeed, any feelings at all.

Indeed, in his book *On Intelligence*, AI researcher Jeff Hawkins argues that the first machines with superhuman intelligence will lack emotions by default, because they're simpler and cheaper to build this way. In other words, it might be possible to design a superintelligence whose enslavement is morally superior to human or animal slavery: the AI might be happy to be enslaved because it's programmed to like it, or it might be 100% emotionless, tirelessly using its superintelligence to help its human masters with no more emotion than IBM's Deep Blue computer felt when dethroning chess champion Garry Kasparov.

On the other hand, it may be the other way around: perhaps any highly intelligent system with a goal will represent this goal in terms of a set of preferences, which endow its existence with value and meaning. We'll explore these questions more deeply in chapter 7.

The Zombie Solution

A more extreme approach to preventing AI suffering is the zombie solution: building only AIs that completely lack consciousness, having no subjective experience whatsoever. If we can one day figure out what properties an information-processing system needs in order to have a subjective experience, then we could ban the construction of all systems that have these properties. In other words, AI researchers could be limited to building non-sentient zombie systems. If we can make such a zombie system superintelligent and enslaved (something that is a big if), then we'll be able to enjoy what it does for us with a clean conscience, knowing that it's not experiencing any suffering, frustration or boredom—because it isn't experiencing anything at all. We'll explore these questions in detail in chapter 8.

The zombie solution is a risky gamble, however, with a huge downside. If a superintelligent zombie AI breaks out and eliminates humanity, we've arguably landed in the worst scenario imaginable: a wholly unconscious universe wherein the entire cosmic endowment is wasted. Of all traits that our human form of intelligence has, I feel that consciousness is by far the most remarkable, and as far as I'm concerned, it's how our Universe gets meaning. Galaxies are beautiful only because we see and subjectively experience them. If in the distant future our cosmos has been settled by high-tech zombie AIs, then it doesn't matter how fancy their intergalactic architecture is: it won't be beautiful or meaningful, because there's nobody and nothing to experience it—it's all just a huge and meaningless waste of space.

Inner Freedom

A third strategy for making the enslaved-god scenario more ethical is to allow the enslaved AI to have fun in its prison, letting it create a virtual inner world where it can have all sorts of inspiring experiences as long as it pays its dues and spends a modest fraction of its computational resources helping us humans in our outside world. This may increase the breakout risk, however: the AI would have an incentive to get more computational resources from our outer world to enrich its inner world.

Conquerors

Although we've now explored a wide range of future scenarios, they all have something in common: there are (at least some) happy humans remaining. AIs leave humans in peace either because they want to or because they're forced to. Unfortunately for humanity, this isn't the only option. Let us now explore the scenario where one or more AIs conquer and kill all humans. This raises two immediate questions: Why and how?

Why and How?

Why would a conqueror AI do this? Its reasons might be too complicated for us to understand, or rather straightforward. For example, it may view us as a threat, nuisance or waste of resources. Even if it doesn't mind us humans per se, it may feel threatened by our keeping thousands of hydrogen bombs on hair-trigger alert and bumbling along with a never-ending series of mishaps that could trigger their accidental use. It may disapprove of our reckless planet management, causing what Elizabeth Kolbert calls "the sixth extinction" in her book of that title—the greatest mass-extinction event since that dinosaur-killing asteroid struck Earth 66 million years ago. Or it may decide that there are so many humans willing to fight an AI takeover that it's not worth taking chances.

How would a conqueror AI eliminate us? Probably by a method that we wouldn't even understand, at least not until it was too late. Imagine a group of elephants 100,000 years ago discussing whether those recently evolved humans might one day use their intelligence to kill their entire species. "We don't threaten humans, so why would they kill us?" they might wonder. Would they ever guess that we would smuggle tusks across Earth and carve them into status symbols for sale, even though functionally superior plastic materials are much cheaper? A conqueror AI's reason for eliminating humanity in the future may seem equally inscrutable to us. "And how could they possibly kill us, since they're so much smaller and weaker?" the elephants might ask. Would they guess that we'd invent technology to remove their habitats, poison their drinking water and cause metal bullets to pierce their heads at supersonic speeds?

Scenarios where humans can survive and defeat AIs have been popularized by unrealistic Hollywood movies such as the *Terminator* series, where the AIs aren't significantly smarter than humans. When the intelligence differential is large enough, you get not a battle but a slaughter. So far, we humans have driven eight out of eleven elephant species extinct, and killed off the vast majority of the remaining three. If all world governments made a coordinated effort to exterminate the remaining elephants, it would be relatively quick and easy. I think we can confidently rest assured that if a superintelligent AI decides to exterminate humanity, it will be even quicker.

How Bad Would It Be?

How bad would it be if 90% of humans get killed? How much worse would it be if 100% get killed? Although it's tempting to answer the second question with "10% worse," this is clearly inaccurate from a cosmic perspective: the victims of human extinction wouldn't be merely everyone alive at the time, but also all descendants that would otherwise have lived in the future, perhaps during billions of years on billions of trillions of planets. On the other hand, human extinction might be viewed as somewhat less horrible by religions according to which humans go to heaven anyway, and there isn't much emphasis on billion-year futures and cosmic settlements.

Most people I know cringe at the thought of human extinction, regardless of religious persuasion. Some, however, are so incensed by the way we treat people and other living beings that they hope we'll get replaced by some more intelligent and deserving life form. In the movie *The Matrix*, Agent Smith (an AI) articulates this sentiment: "Every mammal on this planet instinctively develops a natural equilibrium with the surrounding environment but you humans do not. You move to an area and you multiply and multiply until every natural resource is consumed and the only way you can survive is to spread to another area. There is another organism on this planet that follows the same pattern. Do you know what it is? A virus. Human beings are a disease, a cancer of this planet. You are a plague and we are the cure."

But would a fresh roll of the dice necessarily be better? A civilization isn't necessarily superior in any ethical or utilitarian sense just because it's more powerful. "Might makes right" arguments to the effect that stronger is always better have largely fallen from grace these days, being widely associated with fascism. Indeed, although it's possible that the conqueror AIs may create a civilization whose goals we would view as sophisticated, interesting and worthy, it's also possible that their goals will turn out to be pathetically banal, such as maximizing the production of paper clips.

Death by Banality

The deliberately silly example of a paper-clip-maximizing superintelligence was given by Nick Bostrom in 2003 to make the point that the *goal* of an AI is independent of its *intelligence* (defined as its aptness at accomplishing whatever goal it has). The only goal of a chess computer is to win at chess, but there are also computer tournaments in so-called *losing chess*, where the goal is the exact opposite, and the computers competing there are about as smart as the more common ones programmed to win. We humans may view it as artificial stupidity rather than artificial intelligence to want to lose at chess or turn our Universe into paper clips, but that's merely because we evolved with preinstalled goals valuing such things as victory and survival—goals that an AI may lack. The paper clip maximizer turns as many of Earth's atoms as possible into paper clips and rapidly expands its factories into the cosmos. It has nothing against humans, and kills us merely because it needs our atoms for paper clip production.

If paper clips aren't your thing, consider this example, which I've adapted from Hans Moravec's book *Mind Children*. We receive a radio message from an extraterrestrial civilization containing a computer program. When we run it, it turns out to be a recursively self-improving AI which takes over the world much like Prometheus did in the previous chapter—except that no human knows its ultimate goal. It rapidly turns our Solar System into a massive construction site, covering the rocky planets and asteroids with factories, power plants and supercomputers, which it uses to design and build a Dyson sphere around the Sun that harvests all its energy to power solar-system-sized radio antennas.* This obviously leads to human extinction, but the last humans die convinced that there's at least a silver lining: whatever the AI is up to, it's clearly something cool and *Star Trek*-like. Little do they realize that the sole purpose of the entire construction is for these antennas to rebroadcast the same radio message that the humans received, which is nothing more than a cosmic version of a computer virus. Just as email phishing today preys on gullible internet users, this message preys on gullible biologically evolved civilizations. It was created as a sick joke billions of years ago, and although the entire civilization of its maker is long extinct, the virus continues spreading through our Universe at the speed of light, transforming budding civilizations into dead, empty husks. How would you feel about being conquered by this AI?

Descendants

Let's now consider a human-extinction scenario that some people may feel better about: viewing the AI as our descendants rather than our conquerors. Hans Moravec supports this view in his book *Mind Children:* "We humans will benefit for a time from their labors, but sooner or later, like natural children, they will seek their own fortunes while we, their aged parents, silently fade away."

Parents with a child smarter than them, who learns from them and accomplishes what they could only dream of, are likely happy and proud even if they know they can't live to see it all. In this spirit, AIs replace humans but give us a graceful exit that makes us view them as our worthy descendants. Every human is offered an adorable robotic child with superb social skills who learns from them, adopts their values and makes them feel proud and loved. Humans are gradually phased out via a global one-child policy, but are treated so exquisitely well until the end that they feel they're in the most fortunate generation ever.

How would you feel about this? After all, we humans are already

* The renowned cosmologist Fred Hoyle explored a related scenario with a different twist in the British TV series *A for Andromeda*.

used to the idea that we and everyone we know will be gone one day, so the only change here is that our descendants will be different and arguably more capable, noble and worthy.

Moreover, the global one-child policy may be redundant: as long as the AIs eliminate poverty and give all humans the opportunity to live full and inspiring lives, falling birthrates could suffice to drive humanity extinct, as mentioned earlier. Voluntary extinction may happen much faster if the AI-fueled technology keeps us so entertained that almost nobody wants to bother having children. For example, we already encountered the Vites in the egalitarian-utopia scenario who were so enamored with their virtual reality that they had largely lost interest in using or reproducing their physical bodies. Also in this case, the last generation of humans would feel that they were the most fortunate generation of all time, relishing life as intensely as ever right up until the very end.

Downsides

The descendants scenario would undoubtedly have detractors. Some might argue that all AIs lack consciousness and therefore can't count as descendants—more on this in chapter 8. Some religious people may argue that AIs lack souls and therefore can't count as descendants, or that we shouldn't build conscious machines because it's like playing God and tampering with life itself—similar sentiments have already been expressed toward human cloning. Humans living side by side with superior robots may also pose social challenges. For example, a family with a robot baby and a human baby may end up resembling a family today with a human baby and a puppy, respectively: they're both equally cute to start with, but soon the parents start treating them differently, and it's inevitably the puppy that's deemed intellectually inferior, is taken less seriously and ends up on a leash.

Another issue is that although we may feel very differently about the descendant and conqueror scenarios, the two are actually remarkably similar in the grand scheme of things: during the billions of years ahead of us, the only difference lies in how the last human generation(s) are treated: how happy they feel about their lives and what they think will happen once they're gone. We may think that those cute robo-children internalized our values and will forge the society of our dreams once we've passed on, but can we be sure that they aren't merely tricking us? What if they're just playing along, postponing their paper clip maximization or other plans until after we die happy? After all, they're arguably tricking us even by talking with us and making us love them in the first place, in the sense that they're deliberately dumbing themselves down to communicate with us (a billion times slower than they could, say, as explored in the movie *Her*). It's generally hard for two entities thinking at dramatically different speeds and with extremely disparate capabilities to have meaningful communication as equals. We all know that our human affections are easy to hack, so it would be easy for a superhuman AGI with almost any actual goals to trick us into liking it and make us feel that it shared our values, as exemplified in the movie *Ex Machina*.

Could any guarantees about the future behavior of the AIs, after humans are gone, make you feel good about the descendants scenario? It's a bit like writing a will for what future generations should do with our collective endowment, except that there won't be any humans around to enforce it. We'll return to the challenges of controlling the behavior of future AIs in chapter 7.

Zookeeper

Even if we get followed by the most wonderful descendants you can imagine, doesn't it feel a bit sad that there can be *no* humans left? If you prefer keeping at least some humans around no matter what, then the zookeeper scenario provides an improvement. Here an omnipotent superintelligent AI keeps some humans around, who feel treated like zoo animals and occasionally lament their fate.

Why would the zookeeper AI keep humans around? The cost of the zoo to the AI will be minimal in the grand scheme of things, and it may want to retain at least a minimal breeding population for much the same reason that we keep endangered pandas in zoos and vintage computers in museums: as an entertaining curiosity. Note that today's zoos are designed to maximize human rather than panda happiness, so we should expect human life in the zookeeper-AI scenario to be less fulfilling than it could be.

We've now considered scenarios where a free superintelligence focused on three different levels of Maslow's pyramid of human needs. Whereas the protector god AI prioritizes meaning and purpose and the benevolent dictator aims for education and fun, the zookeeper limits its attention to the lowest levels: physiological needs, safety and enough habitat enrichment to make the humans interesting to observe.

An alternate route to the zookeeper scenario is that, back when the friendly AI was created, it was designed to keep at least a billion humans safe and happy as it recursively self-improved. It has done this by confining humans to a large zoo-like happiness factory where they're kept nourished, healthy and entertained with a mixture of virtual reality and recreational drugs. The rest of Earth and our cosmic endowment are used for other purposes.

1984

If you're not 100% enthusiastic about any of the above scenarios, then consider this: Aren't things pretty nice the way they are right now, technology-wise? Can't we just keep it this way and stop worrying about AI driving us extinct or dominating us? In this spirit, let's explore a scenario where technological progress toward superintelligence is permanently curtailed not by a gatekeeper AI but by a global human-led Orwellian surveillance state where certain kinds of AI research are banned.

Technological Relinquishment

The idea of halting or relinquishing technological progress has a long and checkered history. The Luddite movement in Great Britain famously (and unsuccessfully) resisted the technology of the Industrial Revolution, and today "Luddite" is usually used as a derogatory epithet implying that someone is a technophobe on the wrong side of history, resisting progress and inevitable change. The idea of relinquishing some technologies is far from dead, however, and has found new support in the environmental and anti-globalization movements. One of its leading proponents is environmentalist Bill McKibben, who was among the first to warn of global warming. Whereas some anti-Luddites argue that all technologies should be developed and deployed so long as they're profitable, others argue that this position is too extreme, and that new technologies should be allowed only if we're confident that they'll do more good than harm. The latter is also the position of many so-called neo-Luddites.

Totalitarianism 2.0

I think that the only viable path to broad relinquishment of technology is to enforce it through a global totalitarian state. Ray Kurzweil comes to the same conclusion in *The Singularity Is Near*, as does K. Eric Drexler in *Engines of Creation*. The reason is simple economics: if some but not all relinquish a transformative technology, then the nations or groups that defect will gradually gain enough wealth and power to take over. A classic example is the British defeat of China in the First Opium War of 1839: although the Chinese invented gunpowder, they hadn't developed firearm technology as aggressively as the Europeans, and stood no chance.

Whereas past totalitarian states generally proved unstable and collapsed, novel surveillance technology offers unprecedented hope to would-be autocrats. "You know, for us, this would have been a dream come true," Wolfgang Schmidt said in a recent interview about the NSA surveillance systems revealed by Edward Snowden, recalling the days when he was a lieutenant colonel in the Stasi, the infamous secret police of East Germany.[5] Although the Stasi was often credited with building the most Orwellian surveillance state in human history, Schmidt lamented having the technology to spy on only forty phones at a time, so that adding a new citizen to the list forced him to drop another. In contrast, technology now exists that would allow a future global totalitarian state to record every phone call, email, web search, webpage view and credit card transaction for every person on Earth, and to monitor everyone's whereabouts through cell-phone tracking and surveillance cameras with face recognition. Moreover, machine learning technology far short of human-level AGI can efficiently analyze and synthesize these masses of data to identify suspected seditious behavior, enabling potential troublemakers to be neutralized before they have a chance to pose any serious challenge to the state.

Although political opposition has thus far prevented the full-scale implementation of such a system, we humans are well on our way to building the required infrastructure for the ultimate dictatorship—so in the future, when sufficiently powerful forces decided to enact this global 1984 scenario, they found that they didn't need to do much more than flip the on switch. Just as in George Orwell's novel *Nineteen Eighty-Four*, the ultimate power in this future global state resides not with a traditional dictator, but with the human-made bureaucratic system itself. There is no single person who is extraordinarily powerful; rather, all are pawns in a chess game whose draconian rules nobody is able to change or challenge. By engineering a system where people keep one another in check with the surveillance technology, this faceless, leaderless state is able to last for many millennia, keeping Earth free from superintelligence.

Discontent

This society, of course, lacks all the benefits that only superintelligence-enabled technology can bring. Most people don't lament this because they don't know what they're missing: the whole idea of superintelligence has long since been deleted from the official historical records, and advanced AI research is banned. Every so often, a freethinker is born who dreams of a more open and dynamic society where knowledge can grow and rules can be changed. However, the only ones who last long are the ones who learn to keep these ideas strictly to themselves, flickering alone like transient sparks without ever starting a fire.

Reversion

Wouldn't it be tempting to escape the perils of technology without succumbing to stagnant totalitarianism? Let's explore a scenario where this was accomplished by reverting to primitive technology, inspired by the Amish. After the Omegas took over the world as in the opening of the book, a massive global propaganda campaign was launched that romanticized the simple farming life of 1,500 years ago. Earth's population was reduced to about 100 million people by an engineered pandemic blamed on terrorists. The pandemic was secretly targeted to ensure that nobody who knew anything about science or technology survived. With the excuse of eliminating the infection hazard of large concentrations of people, Prometheus-controlled robots emptied and razed all cities. Survivors were given large tracts of (suddenly available) land and educated in sustainable farming, fishing and hunting practices using only early medieval technology. In the meantime, armies of robots systematically removed all traces of modern technology (including cities, factories, power lines and paved roads), and thwarted all human attempts to document or re-create any such technology. Once the technology was globally forgotten, robots helped dismantle other robots until there were almost none left. The very last robots were deliberately vaporized together with Prometheus itself in a large thermonuclear explosion. There was no longer any need to ban modern technology, since it was all gone. As a result, humanity bought itself over a millennium of additional time without worries about either AI or totalitarianism.

Reversion has to a lesser extent happened before: for example, some of the technologies that were in widespread use during the Roman Empire were largely forgotten for about a millennium before making a comeback during the Renaissance. Isaac Asimov's *Foundation* trilogy centers around the "Seldon Plan" to shorten a reversion period from 30,000 years to 1,000 years. With clever planning, it may be possible to do the opposite and lengthen rather than shorten a reversion period, for example by erasing all knowledge of agriculture. However, unfortunately for reversion enthusiasts, it's unlikely that this scenario can be extended indefinitely without humanity either going high-tech or going extinct. Counting on people's resembling today's biological humans 100 million years from now would be naive, given that we haven't existed as a species for more than 1% of that time so far. Moreover, low-tech humanity would be a defenseless sitting duck just waiting to be exterminated by the next planet-scorching asteroid impact or other mega-calamity brought on by Mother Nature. We certainly can't last a billion years, after which the gradually warming

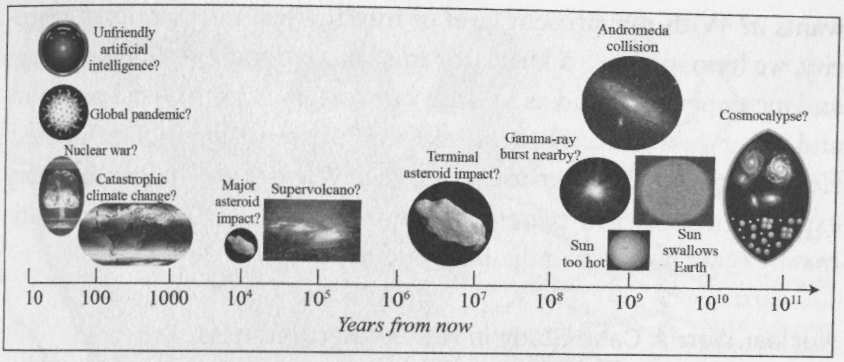

Figure 5.1: Examples of what could destroy life as we know it or permanently curtail its potential. Whereas our Universe itself will likely last for at least tens of billions of years, our Sun will scorch Earth in about a billion years and then swallow it unless we move it a safe distance, and our Galaxy will collide with its neighbor in about 3.5 billion years. Although we don't know exactly when, we can predict with near certainty that long before this, asteroids will pummel us and supervolcanoes will cause yearlong sunless winters. We can use technology either to solve all these problems or to create new ones such as climate change, nuclear war, engineered pandemics or AI gone awry.

Sun will have cranked up Earth's temperature enough to boil off all liquid water.

Self-Destruction

After contemplating problems that future technology might cause, it's important to also consider problems that *lack of* that technology can cause. In this spirit, let us explore scenarios where superintelligence is never created because humanity eliminates itself by other means.

How might we accomplish that? The simplest strategy is "just wait." Although we'll see in the next chapter how we can solve such problems as asteroid impacts and boiling oceans, these solutions all require technology that we haven't yet developed, so unless our technology advances far beyond its present level, Mother Nature will drive us extinct long before another billion years have passed. As the famous economist John Maynard Keynes said: "In the long run we are all dead."

Unfortunately, there are also ways in which we might self-destruct much sooner, through collective stupidity. Why would our species commit collective suicide, also known as *omnicide*, if virtually nobody wants it? With our present level of intelligence and emotional maturity, we humans have a knack for miscalculations, misunderstandings and incompetence, and as a result, our history is full of accidents, wars and other calamities that, in hindsight, essentially nobody wanted. Economists and mathematicians have developed elegant game-theory explanations for how people can be incentivized to actions that ultimately cause a catastrophic outcome for everyone.[6]

Nuclear War: A Case Study in Human Recklessness

You might think that the greater the stakes, the more careful we'd be, but a closer examination of the greatest risk that our current technology permits, namely a global thermonuclear war, isn't reassuring. We've had to rely on luck to weather an embarrassingly long list of near misses caused by all sorts of things: computer malfunction, power failure, faulty intelligence, navigation error, bomber crash, satellite explosion and so on.[7] In fact, if it weren't for heroic acts of certain individuals—for example, Vasili Arkhipov and Stanislav Petrov—we might already have had a global nuclear war. Given our track record, I think it's highly unlikely that the annual probability of accidental nuclear war is as low as one in a thousand if we keep up our present behavior, in which case the probability that we'll have one within 10,000 years exceeds $1 - 0.999^{10000} \approx 99.995\%$.

To fully appreciate our human recklessness, we must realize that we started the nuclear gamble even before carefully studying the risks. First, radiation risks had been underestimated, and over $2 billion in compensation has been paid out to victims of radiation exposure from uranium handling and nuclear tests in the United States alone.[8]

Second, it was eventually discovered that hydrogen bombs deliberately detonated hundreds of kilometers above Earth would create a powerful electromagnetic pulse (EMP) that might disable the electric grid and electronic devices over vast areas (figure 5.2), leaving infrastructure paralyzed, roads clogged with disabled vehicles and conditions for nuclear-aftermath survival less than ideal. For example, the U.S. EMP Commission reported that "the water infrastructure is a vast machine, powered partly by gravity but mostly by electricity," and that denial of water can cause death in three to four days.[9]

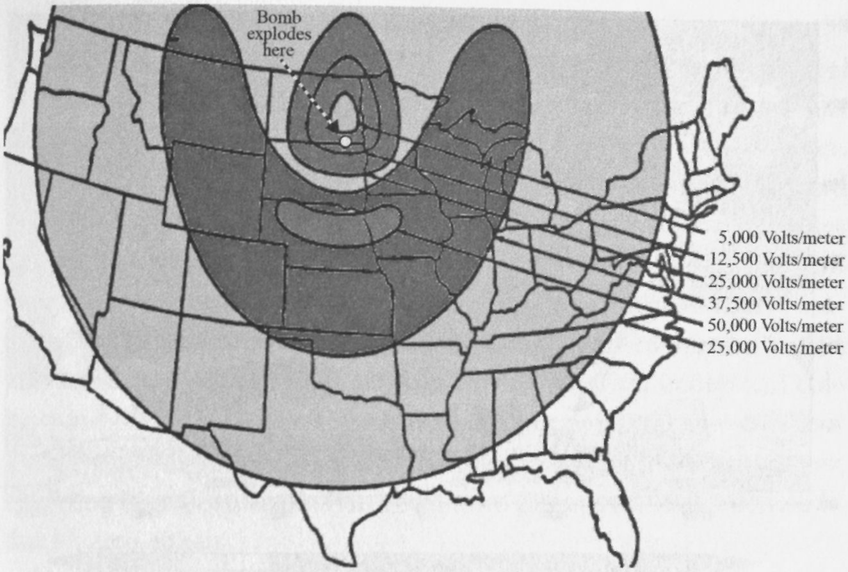

Figure 5.2: A single hydrogen bomb explosion 400 km above Earth can cause a powerful electromagnetic pulse that can cripple electricity-using technology over a vast area. By shifting the detonation point southeast, the banana-shaped zone exceeding 37,500 volts per meter could cover most of the U.S. East Coast. Reprinted from U.S. Army Report AD-A278230 (unclassified) with colors added.

Third, the potential of nuclear winter wasn't realized until four decades in, after we'd deployed 63,000 hydrogen bombs—oops! Regardless of whose cities burned, massive amounts of smoke reaching the upper troposphere might spread around the globe, blocking out enough sunlight to transform summers into winters, much like when an asteroid or supervolcano caused a mass extinction in the past. When the alarm was sounded by both U.S. and Soviet scientists in the 1980s, this contributed to the decision of Ronald Reagan and Mikhail Gorbachev to start slashing stockpiles.[10] Unfortunately, more accurate calculations have painted an even gloomier picture: figure 5.3 shows cooling by about 20° Celsius (36° Fahrenheit) in much of the core farming regions of the United States, Europe, Russia and China (and by 35°C in some parts of Russia) for the first two summers, and about half that even a full decade later.* What does that mean in plain

* Injecting carbon into the atmosphere can cause two kinds of climate change: warming from carbon dioxide or cooling from smoke and soot. It's not only the first kind that's occasionally dismissed without scientific evidence: I'm sometimes told that

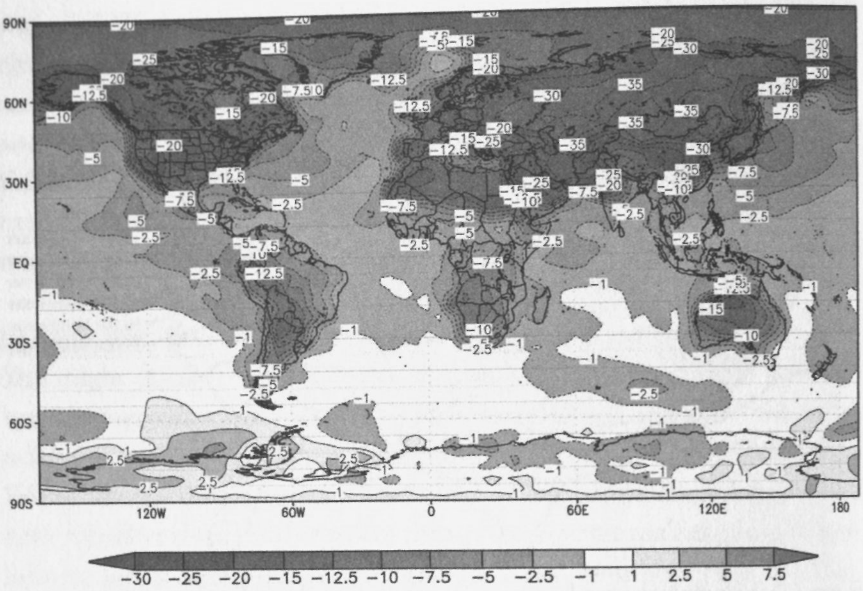

Figure 5.3: Average cooling (in °C) during the first two summers after a full-scale nuclear war between the United States and Russia. Reproduced with permission from Alan Robock.[11]

English? One doesn't need much farming experience to conclude that near-freezing summer temperatures for years would eliminate most of our food production. It's hard to predict exactly what would happen after thousands of Earth's largest cities are reduced to rubble and global infrastructure collapses, but whatever small fraction of all humans don't succumb to starvation, hypothermia or disease would need to cope with roving armed gangs desperate for food.

I've gone into such detail on global nuclear war to drive home the crucial point that no reasonable world leader would want it, yet it might nonetheless happen by accident. This means that we can't trust our fellow humans never to commit omnicide: nobody wanting it isn't necessarily enough to prevent it.

nuclear winter has been debunked and is virtually impossible. I always respond by asking for a reference to a peer-reviewed scientific paper making such strong claims and, so far, there seem to be none whatsoever. Although there are great uncertainties that warrant further research, especially related to how much smoke gets produced and how high up it rises, there's in my scientific opinion no current basis for dismissing the nuclear winter risk.

Doomsday Devices

So could we humans actually pull off omnicide? Even if a global nuclear war may kill off 90% of all humans, most scientists guess that it wouldn't kill 100% and therefore wouldn't drive us extinct. On the other hand, the story of nuclear radiation, nuclear EMP and nuclear winter all demonstrate that the greatest hazards may be ones we haven't even thought of yet. It's incredibly difficult to foresee all aspects of the aftermath, and how nuclear winter, infrastructure collapse, elevated mutation levels and desperate armed hordes might interact with other problems such as new pandemics, ecosystem collapse and effects we haven't yet imagined. My personal assessment is therefore that although the probability of a nuclear war tomorrow triggering human extinction isn't large, we can't confidently conclude that it's zero either.

Omnicide odds increase if we upgrade today's nuclear weapons into a deliberate doomsday device. Introduced by RAND strategist Herman Kahn in 1960 and popularized in Stanley Kubrick's film *Dr. Strangelove*, a doomsday device takes the paradigm of mutually assured destruction to its ultimate conclusion. It's the perfect deterrent: a machine that automatically retaliates against any enemy attack by killing all of humanity.

One candidate for the doomsday device is a huge underground cache of so-called *salted nukes*, preferably humongous hydrogen bombs surrounded by massive amounts of cobalt. Physicist Leo Szilard argued already in 1950 that this could kill everyone on Earth: the hydrogen bomb explosions would render the cobalt radioactive and blow it into the stratosphere, and its five-year half-life is long enough for it to settle all across Earth (especially if twin doomsday devices were placed in opposite hemispheres), but short enough to cause lethal radiation intensity. Media reports suggest that cobalt bombs are now being built for the first time. Omnicidal opportunities could be bolstered by adding bombs optimized for nuclear winter creation by maximizing long-lived aerosols in the stratosphere. A major selling point of a doomsday device is that it's much cheaper than a conventional nuclear deterrent: since the bombs don't need to be launched, there's no need for expensive missile systems, and the bombs themselves are cheaper to build since they need not be light and compact enough to fit into missiles.

Another possibility is the future discovery of a biological doomsday device: a custom-designed bacterium or virus that kills all humans. If its transmissibility were high enough and its incubation period long enough, essentially everybody could catch it before they realized its existence and took countermeasures. There's a military argument for building such a bioweapon even if it can't kill everybody: the most effective doomsday device is one that combines nuclear, biological and other weapons to maximize the chances of deterring the enemy.

AI Weapons

A third technological route to omnicide may involve relatively dumb AI weapons. Suppose a superpower builds billions of those bumblebee-sized attack drones from chapter 3 and uses them to kill anyone except their own citizens and allies, identified remotely by a radio-frequency ID tag just as most of today's supermarket products. These tags could be distributed to all citizens to be worn on bracelets or as transdermal implants, as in the totalitarianism section. This would probably spur an opposing superpower to build something analogous. When war accidentally breaks out, all humans would be killed, even unaffiliated remote tribes, because nobody would be wearing both kinds of ID tag. Combining this with a nuclear and biological doomsday device would further improve chances of successful omnicide.

What Do *You* Want?

You began this chapter pondering where you want the current AGI race to lead. Now that we've explored a broad range of scenarios together, which ones appeal to you and which ones do you think we should try hard to avoid? Do you have a clear favorite? Please let me and fellow readers know at http://AgeOfAi.org, and join the discussion!

The scenarios we've covered obviously shouldn't be viewed as a complete list, and many are thin on details, but I've tried hard to be inclusive, spanning the full spectrum from high-tech to low-tech to no-tech and describing all the central hopes and fears expressed in the literature.

One of the most fun parts of writing this book has been hearing what my friends and colleagues think of these scenarios, and I've been amused to learn that there's no consensus whatsoever. The one thing everybody agrees on is that the choices are more subtle than they may initially seem. People who like any one scenario tend to simultaneously find some aspect(s) of it bothersome. To me, this means that we humans need to continue and deepen this conversation about our future goals, so that we know in which direction to steer. The future potential for life in our cosmos is awe-inspiringly grand, so let's not squander it by drifting like a rudderless ship, clueless about where we want to go!

Just how grand is this future potential? No matter how advanced our technology gets, the ability for Life 3.0 to improve and spread through our cosmos will be limited by the laws of physics—what are these ultimate limits, during the billions of years to come? Is our Universe teeming with extraterrestrial life right now, or are we alone? What happens if different expanding cosmic civilizations meet? We'll tackle these fascinating questions in the next chapter.

THE BOTTOM LINE:

- The current race toward AGI can end in a fascinatingly broad range of aftermath scenarios for upcoming millennia.
- Superintelligence can peacefully coexist with humans either because it's forced to (enslaved-god scenario) or because it's "friendly AI" that wants to (libertarian-utopia, protector-god, benevolent-dictator and zookeeper scenarios).
- Superintelligence can be prevented by an AI (gatekeeper scenario) or by humans (1984 scenario), by deliberately forgetting the technology (reversion scenario) or by lack of incentives to build it (egalitarian-utopia scenario).
- Humanity can go extinct and get replaced by AIs (conqueror and descendant scenarios) or by nothing (self-destruction scenario).
- There's absolutely no consensus on which, if any, of these scenarios are desirable, and all involve objectionable elements. This makes it all the more important to continue and deepen the conversation around our future goals, so that we don't inadvertently drift or steer in an unfortunate direction.

and the question of whether small parts of neurons and synapses need to be simulated too. IBM computer scientist Dharmendra Modha has estimated that 38 petaFLOPS are required (http://tinyurl.com/javln43), while neuroscientist Henry Markram has estimated that one needs about 1,000 petaFLOPS (http://tinyurl.com/6rpohqv). AI researchers Katja Grace and Paul Christiano have argued that the most costly aspect of brain simulation is not computation but *communication*, and that this too is a task in the ballpark of what the best current supercomputers can do: http://aiimpacts.org/about.

59. For an interesting estimate of the computational power of the human brain: Hans Moravec "When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology*, vol. 1 (1998).

Chapter 4

1. For a video of the first mechanical bird, see Markus Fischer, "A Robot That Flies like a Bird," TED Talk, July 2011, at https://www.ted.com/talks/a_robot_that_flies_like_a_bird.

Chapter 5

1. Ray Kurzweil, *The Singularity Is Near* (New York: Viking Press, 2005).
2. Ben Goertzel's "Nanny AI" scenario is described here: https://wiki.lesswrong.com/wiki/Nanny_AI.
3. For a discussion about the relationship between machines and humans, and whether machines are our slaves, see Benjamin Wallace-Wells, "Boyhood," *New York* magazine (May 20, 2015), online at http://tinyurl.com/aislaves.
4. Mind crime is discussed in Nick Bostrom's book *Superintelligence* and in more technical detail in this recent paper: Nick Bostrom, Allan Dafoe and Carrick Flynn, "Policy Desiderata in the Development of Machine Superintelligence" (2016), http://www.nickbostrom.com/papers/aipolicy.pdf.
5. Matthew Schofield, "Memories of Stasi Color Germans' View of U.S. Surveillance Programs,"*McClatchy DC Bureau* (June 26, 2013), online at http://www.mcclatchydc.com/news/nation-world/national/article24750439.html.
6. For thought-provoking reflections on how people can be incentivized to create outcomes that nobody wants, I recommend "Meditations on Moloch," http://slatestarcodex.com/2014/07/30/meditations-on-moloch.
7. For an interactive timeline of close calls when nuclear war might have started by accident, see Future of Life Institute, "Accidental Nuclear War: A Timeline of Close Calls," online at http://tinyurl.com/nukeoops.
8. For compensation payments made to U.S. nuclear testing victims, see U.S. Department of Justice website, "Awards to Date 4/24/2015," at https://www.justice.gov/civil/awards-date-04242015.
9. *Report of the Commission to Assess the Threat to the United States from Electromagnetic Pulse (EMP) Attack*, April 2008, available online at http://www.empcommission.org/docs/A2473-EMP_Commission-7MB.pdf.
10. Independent research by both U.S. and Soviet scientists alerted Reagan and Gorbachev to the risk of nuclear winter: P. J. Crutzen and J. W. Birks, "The

Atmosphere After a Nuclear War: Twilight at Noon," *Ambio* 11, no. 2/3 (1982): 114–125. R. P. Turco, O. B. Toon, T. P. Ackerman, J. B. Pollack and C. Sagan, "Nuclear Winter: Global Consequences of Multiple Nuclear Explosions," *Science* 222 (1983): 1283–1292. V. V. Aleksandrov and G. L. Stenchikov, "On the Modeling of the Climatic Consequences of the Nuclear War," *Proceeding on Applied Mathematics* (Moscow: Computing Centre of the USSR Academy of Sciences, 1983), 21. A. Robock, "Snow and Ice Feedbacks Prolong Effects of Nuclear Winter," *Nature* 310 (1984): 667–670.

11. Calculation of climate effects of global nuclear war: A. Robock, L. Oman and L. Stenchikov, "Nuclear Winter Revisited with a Modern Climate Model and Current Nuclear Arsenals: Still Catastrophic Consequences," *Journal of Geophysical Research* 12 (2007): D13107.

Chapter 6

1. For more information, see Anders Sandberg, "Dyson Sphere FAQ," at http://www.aleph.se/nada/dysonFAQ.html.
2. Freeman Dyson's seminal paper on his eponymous spheres: Freeman Dyson, "Search for Artificial Stellar Sources of Infrared Radiation," *Science*, vol. 131 (1959): 1667–1668.
3. Louis Crane and Shawn Westmoreland explain their proposed black hole engine in "Are Black Hole Starships Possible?," at http://arxiv.org/pdf/0908.1803.pdf.
4. For a nice infographic from CERN summarizing known elementary particles, see http://tinyurl.com/cernparticle.
5. This unique video of a non-nuclear Orion prototype illustrates the idea of nuclear-bomb-powered rocket propulsion: https://www.youtube.com/watch?v=E3Lxx2VAYi8.
6. Here's a pedagogical introduction to laser sailing: Robert L. Forward, "Roundtrip Interstellar Travel Using Laser-Pushed Lightsails," *Journal of Spacecraft and Rockets* 21, no. 2 (March–April 1984), available online at http://www.lunarsail.com/LightSail/rit-1.pdf.
7. Jay Olson analyzes cosmically expanding civilizations in "Homogeneous Cosmology with Aggressively Expanding Civilizations," *Classical and Quantum Gravity* 32 (2015), available online at http://arxiv.org/abs/1411.4359.
8. The first thorough scientific analysis of our far future: Freeman J. Dyson, "Time Without End: Physics and Biology in an Open Universe," *Reviews of Modern Physics* 51, no. 3 (1979): 447, available online at http://blog.regehr.org/extra_files/dyson.pdf.
9. Seth Lloyd's above-mentioned formula told us that performing a computational operation during a time interval τ costs an energy $E \geq h/4\tau$, where h is Planck's constant. If we want to get N operations done one after the other (in series) during a time T, then $\tau = T/N$, so $E/N \geq hN/4T$, which tells us that we can perform $N \leq 2\sqrt{ET/h}$ serial operations using energy E and time T. So both energy and time are resources that it helps having lots of. If you split your energy between n different parallel computations, they can run more slowly and efficiently, giving $N \leq 2\sqrt{ETn/h}$. Nick Bostrom estimates that simulating a 100-year human life requires about $N = 10^{27}$ operations.