1

Introduction

Everything was finally in place. The scientists stood in their observation bunker in the Alamogordo desert, staring through the lenses of their protective glasses at the tower five miles away where the world's first atomic device awaited an electric pulse for detonation. It was 5:29 a.m. on July 16, 1945.

The previous evening, one of the smartest persons in the world had decided it was time to lighten things up a bit. Enrico Fermi glanced round at his fellow scientists and said, in his heavy Italian accent, "Now, let's make a bet whether the atmosphere will be set on fire by this test."[1] He wondered aloud whether this would incinerate only New Mexico or spread to the entire planet. Laughter followed, some of it nervous. A few people took him up on it.

In actual fact, they had seriously examined this possibility three years earlier. The Nobel laureate Hans Bethe later recalled how it went.

> BETHE: So one day at Berkeley – we were a very small group, maybe eight physicists or so – one day Teller came to the office and said, "Well, what would happen to the air if an atomic bomb were exploded in the air? . . . There's nitrogen in the air, and you can have a nuclear reaction in which two nitrogen nuclei collide and become oxygen plus carbon, and in this process you set free a lot of energy. Couldn't that happen?" And that caused great excitement.
>
> INTERVIEWER: This is in '42?
>
> BETHE: '42. Oppenheimer got quite excited and said, "That's a terrible possibility," and he went to his superior, who was Arthur Compton, the director of the Chicago Laboratory, and told him that. Well, I sat down and looked at the problem, about whether two nitrogen nuclei could penetrate each other and make that nuclear reaction, and I found that it was just incredibly unlikely . . . Teller at Los Alamos put a very good calculator on this problem, [Emil] Konopinski, who was an expert on weak interactors, and Konopinski together with [inaudible] showed that it

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

was incredibly impossible to set the hydrogen, to set the atmosphere on fire. They wrote one or two very good papers on it, and that put the question really at rest. They showed in great detail why it is impossible.

INTERVIEWER: I think what makes it such a fascinating episode . . . is the idea of doing a calculation on which possibly could rest the fate of the world. [laughter]

BETHE: Right, right.

INTERVIEWER: That's obviously an extraordinary kind of calculation to do. Did you have any . . . Did you even think about that issue when you saw the Trinity test?

BETHE: No.

INTERVIEWER: You were absolutely –

BETHE: Yes.

INTERVIEWER: – completely certain.

BETHE: Yes. The one thing in my mind was that maybe the initiator would not work because I had a lot to do with its design, and that the whole thing would be a fizzle because the initiator wasn't working. No, it never occurred to me that it would set the atmosphere on fire.

INTERVIEWER: In a way, this is like a great test of one's belief –

BETHE: In science. [laughter][2]

The firing sequence for the atomic test was automated, but four individuals along the chain of command had been given kill switches to activate if anything went wrong. The final-stage knife switch rested in the hands of a 25-year-old chemist named Donald Hornig:

> I don't think I have ever been keyed up as I was during those final seconds . . . I kept telling myself "the least flicker of that needle and you have to act." It kept on coming down to zero. I kept saying, "Your reaction time is about half a second and you can't relax for even a fraction of a second." . . . My eyes were glued on the dial and my hand was on the switch. I could hear the timer counting . . . three . . . two . . . one. The needle fell to zero.[3]

The Harvard scientist James Conant, one of the Manhattan Project's overseers, squinted through his protective glasses from the nearby bunker.

> Then came a burst of white light that seemed to fill the sky and seemed to last for seconds. I had expected a relatively quick and bright flash. The enormity of the light quite stunned me. My instantaneous reaction was that something had gone wrong and that the thermal nuclear transformation of the atmosphere, once discussed as a possibility and jokingly referred to minutes earlier, had actually occurred.[4]

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

The international team of scientists had also taken bets among themselves about what sort of bang the explosion would yield. No one had ever built a device like this before, so they couldn't be sure. The bets ranged from zero (a dud) to 45 kilotons (a kiloton is equivalent to two million pounds of TNT). Isidor Rabi, a physicist who arrived late for the betting pool, took the last remaining option by default: 18 kilotons. He won. The actual Trinity explosion weighed in at 22.1 kilotons.

PLANETARY DANGERS, PLANETARY STRATEGIES

Accidentally setting the atmosphere on fire: this is a book about the special kind of wager we are making when we create technologies that put the survival of humankind at risk. Four kinds of inventions fall into this category, and although they each spring from very different arenas of scientific research, they share similar properties: they are radical game-changers, exceptionally powerful at doing what they do; they have the potential to bring tremendous benefit as well as harm into human affairs; and they are devilishly hard to control, in two distinct ways. It's difficult to guarantee their proper functioning, from a technical point of view; and it's even harder to govern the political, economic, and military processes through which they are brought into the world and put to use. Their awesome power makes it hard to resist pursuing them; yet if we handle them improperly they may destroy us all.

The first of these, fossil-fuel-based technology, dates back 200 years to the early 1800s: it was one of the factors that launched the Industrial Revolution. The second, nuclear technology, was born on that July morning near the end of World War II. The other two are more recent inventions, still far from realizing the full scope of their promise. Synthetic biology aims to make new life forms from scratch, creating microorganisms that serve human welfare in unprecedented ways – yet it also presents us with the possibility of hideous bioweapons or bioengineered pandemics. Artificial intelligence (AI) seeks to endow our machines with one of the most valuable qualities any creature possesses: the ability to figure things out for itself and bring genuine innovation into the world. Yet if advanced AI escapes human control, or is weaponized for military use, it too could bring down catastrophic consequences on its makers.

These four technologies create planet-level dangers that can only be addressed effectively with planet-level solutions. Precisely because they are so powerful, human beings compete with each other to obtain them, racing

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

frantically to develop them to ever more impressive levels of potency. And wherever the logic of high-stakes competition prevails, the logics of caution, transparency, and cooperation tend to take a back seat. Despite the grievous risks inherent in these inventions, efforts to regulate or rein in their development have failed, under the same recurring refrain: "If we don't build these things and tap their power, someone else will beat us to it and we'll be at the mercy of the winners. So damn the torpedoes and full steam ahead!"

In this book I survey the promise and dangers of these four technologies, as well as the strategies proposed by experts in each field for bringing them under better systems of control. The main brunt of my argument focuses on the international dimension, where every sensible solution tends to fail because of the dog-eat-dog competition that prevails in the global arena. I devote a lot of thought to this problem because it tends to be the aspect that gets shortest shrift in most other discussions about controlling mega-dangerous technologies. The reason is simple: most authors tacitly assume that the present-day system of international relations – separate sovereign nations fending for themselves in the win–lose game of geopolitics – will continue to hold sway over the coming centuries. My analysis takes a quite different tack. I maintain that the separateness of nations is already being blurred in fundamental ways by the economic and technological forces of globalization: we live in a world of deepening interdependence and shared vulnerabilities. Operating on this premise, I conclude that our best shot at tackling these planet-scale problems is to gradually build a comprehensive framework of global cooperation and collective security over the next century.

In much of the expert literature, creating such a framework tends to be dismissed as a utopian fantasy. "Maybe a millennium from now, if we're very lucky" – this is the assumption. But building a better global governance system is actually quite feasible if we put our minds to it over the coming decades. I argue that we are already farther along in this process than most people realize, and that although the road ahead is long and difficult, this is a challenge that humankind is well equipped to take on successfully.

The key to my approach lies in the idea of purposeful, incremental change. Some of the most profound transformations of the modern era have come about gradually, through the dogged efforts of successive generations of committed individuals. The realm of the possible is not a stable place: it changes from decade to decade, as people's habits, expectations, and assumptions evolve. Technological capabilities that seemed fanciful two centuries ago – such as conversing casually with someone on the other side of the planet – became conceivable a century ago, then a reality 50 years ago.

Societal achievements deemed utopian in 1850 – such as true equality for women – became legally the norm in recent decades and are on their way to becoming a socioeconomic reality in the coming years. In 1950 the Western European nations were reeling from the latest round of blood-drenched conflict that had just convulsed their continent, following centuries of rivalry, wars, and mutual distrust; yet four decades later these same nations were signing the Maastricht Treaty, binding themselves together in a partial supranational union that rendered war among them about as likely as armed aggression between the United States and Canada. Small, purpose-driven changes are like compound interest in this sense: they can accumulate powerfully over time, yielding quietly revolutionary results.

Some would argue that the effort to envision a full-fledged system of global governance is a waste of time, and that we are better served by focusing on tangible measures that humankind can feasibly adopt within the constraints of today's international system. My response is that it's helpful to do *both* these things: to work out the key practical dimensions of the long-term goal, while also offering intermediate steps that can help us mitigate these mega-dangers today. Over the coming decades, we are more likely to make meaningful progress toward cooperative security if we have a relatively clear sense of the long-range principles and institutional arrangements we will ultimately need to realize.

This is how societal innovation emerges.[5] A few visionary individuals push the limits of normal behavior, eliciting frowns of surprise and disapproval from their contemporaries; time passes, and other plucky individuals join in; eventually more and more people climb on the bandwagon, adding to the critical mass of the disruptive trend; and gradually the novel behavior pattern stabilizes and becomes the new normal. (Even the original frowners usually join in at this point, claiming to have been on board all along.) After a while a different group of boundary-pushers invents yet another trailblazing behavior to go beyond norms in their own original way. Generation by generation, the alterations accumulate: problems that once baffled people are easily solved; institutions that were deemed utopian come unobtrusively into being; achievements that seemed fanciful to one's grandparents are now a daily routine.

At present, the idea of a vigorous, democratic United Nations (UN) coordinating the governance of the world's affairs strikes most people as a distant dream. But with each international crisis that humankind confronts and works through – with each planet-level problem that forces people to come up with new planet-level solutions – the boundaries of the possible shift slightly. To someone situated in the year 2075, a revitalized UN

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

may no longer seem nearly as implausible as it does today; and to a citizen of the year 2150, many aspects of it may already be taken for granted as an accomplished fact.

As a historian, however, I know that human beings have a knack for not always doing what's in their best interest. It's quite possible that a deep political transformation will elude us over the coming decades, and that we'll reach the year 2150 saddled with roughly the same ramshackle international mechanisms operating today. Yet even this more muddled and precarious pathway can be trodden in relatively smart or stupid ways. There remain plenty of concrete interim solutions we can undertake to buy ourselves additional precious time. It may turn out that our species lacks the vision or the will to create an effective global government over the coming century or so; but we can still adopt plenty of sensible precautionary measures that boost our long-term chances of making it through in one piece.

UNCONTROLLED EXPONENTIAL PROCESSES

All four of these technologies exhibit an intriguing common feature: the potential for uncontrolled exponential growth. This quality sets them apart from more conventional technologies and lies at the root of what makes them so dangerous. The thermostat in my house is an autonomous machine, operating on its own to manage the ambient temperature; but it poses no big risk because it has no exponential properties in its functioning. A bacterial culture in a petri dish follows an exponential trajectory as its cells replicate and grow; but it's not problematic because its propagation is constrained by the nutrients provided by its human overseers. The Spanish flu virus of 1918–20, by contrast, exhibited strong elements of both exponential and autonomous functioning, and was limited only by the number of hosts it could contact and infect: this is why it killed 50 million people worldwide (far more than the carnage of World War I).

At first glance, fossil fuels would seem not to fit this description, since there is nothing inherently exponential about the way they work – but the *history* of their development does reveal an unmistakably exponential trajectory (see Figure 1.1).

People started using coal on a large scale in the mid-1800s, and the tapping of this novel power source helped incentivize the invention of a wide variety of new machines, whose rapid propagation in turn facilitated the accelerating development of petroleum and natural gas. Technological advance and energy innovation fed off each other in a self-reinforcing

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

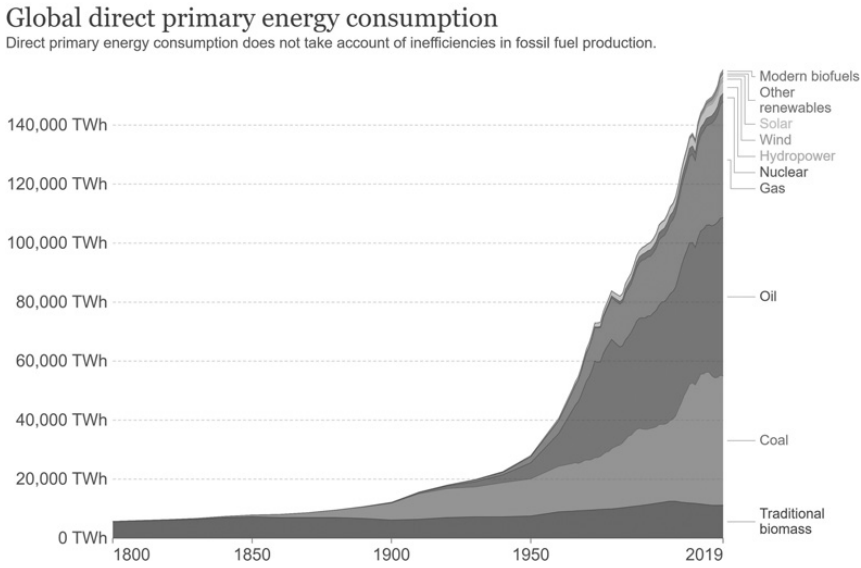

Figure 1.1 Global primary energy consumption, 1800–2019. The net growth of global energy consumption in the ten years between 2000 and 2010 (+35,000 terawatt hours) was the same as the net growth in the *50* years between 1910 and 1960. A terawatt hour is one trillion watt hours, about the same as the output of 1,600 of today's coal-fired power plants.[6]
Source: Vaclav Smil (2017) & BP Statistical Review of World Energy

spiral, and their development traced a steeply rising curve that grew dramatically over the course of the following century, with a particularly sharp inflection in the decades since 1945.[7] Much like the influenza of 1918, this exponentially rising process was limited only by the total quantity of fossil fuel resources available on the planet.

In a similar way, the chain reaction at the heart of nuclear devices is based on unleashing an exponential propagation of fission or fusion through a fuel medium (see Figure 1.2).

At the first step in the process, the nucleus of an atom of fuel material splits, releasing two neutrons that go on to strike up to two other nearby atoms, causing them to split in turn, and the exponential cycle is off and running: 2, 4, 8, 16, 32, and so on. With each splitting of an atom, a small amount of energy is released in the form of heat. Since the process propagates swiftly, reaching trillions of splitting atoms in a matter of microseconds, this results in a very large amount of heat energy accumulating in a small space. The energy expands outward, appearing to external observers as an explosion. A big one.

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

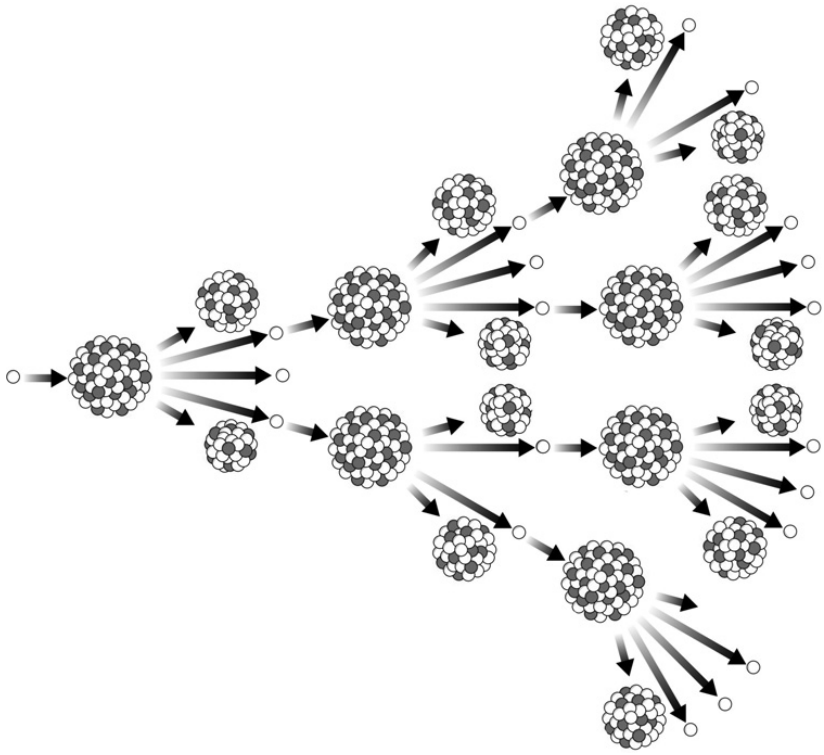

Figure 1.2 The nuclear fission chain reaction.[8]

Some of the most potent forms of synthetic biology operate in a similar fashion to naturally occurring pandemics: they rely on a small number of seed organisms designed to make an exponential number of copies of themselves by using materials from the surrounding environment. (See Figure 1.3.)

Harnessing the reproductive power of engineered microorganisms could allow humans to turn algae into biofuels, fungi into medicines, and toxic garbage into harmless sludge. Ideally, these synthetic life forms will remain limited in their capacity to self-propagate, either because they have been designed to stop replicating after a certain number of reproductive cycles, or because they are constrained by the particular forms of nutrients they have been designed to consume and transform. But if a laboratory error or unforeseen mutation allows them to escape containment – or if they are deliberately unleashed by malevolent actors – their runaway propagation could wreak grievous large-scale harm.

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

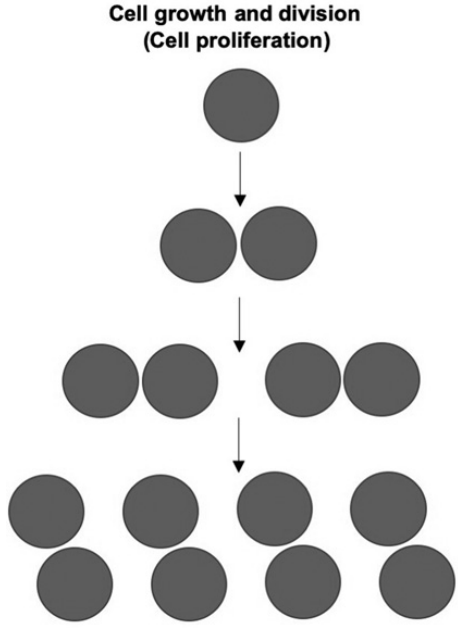

Figure 1.3 Cell growth process.[9]

The most impressive forms of AI will be self-modifying machines that learn about their environment as they interact with it, refining their own software and hardware as they go. Already today, such machines are being designed and built. One game-playing AI built by Google, for example, can translate its pattern-recognition, reinforcement learning, and strategy-development abilities from one game to another to another – without any human guidance or intervention. When presented with a new and unfamiliar game, it engages in increasingly sophisticated cycles of trial-and-error and achieves mastery at the game in a rapid spiral of self-improvement.[10] (See Figure 1.4.)

Today's self-improving machines are narrowly focused on specific tasks, but what will happen over the coming decades, as such machines become increasingly adept at performing a wide variety of different functions? What happens when the AI-controlled robot that drives your car can also trade stocks, play chess, diagnose your illness, assemble furniture, solve math problems, give you relationship advice, cook delicious meals, and pass high-level intelligence tests?[11] Such an all-purpose machine will presumably be able to modify its own software, rendering itself better attuned to achieving

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

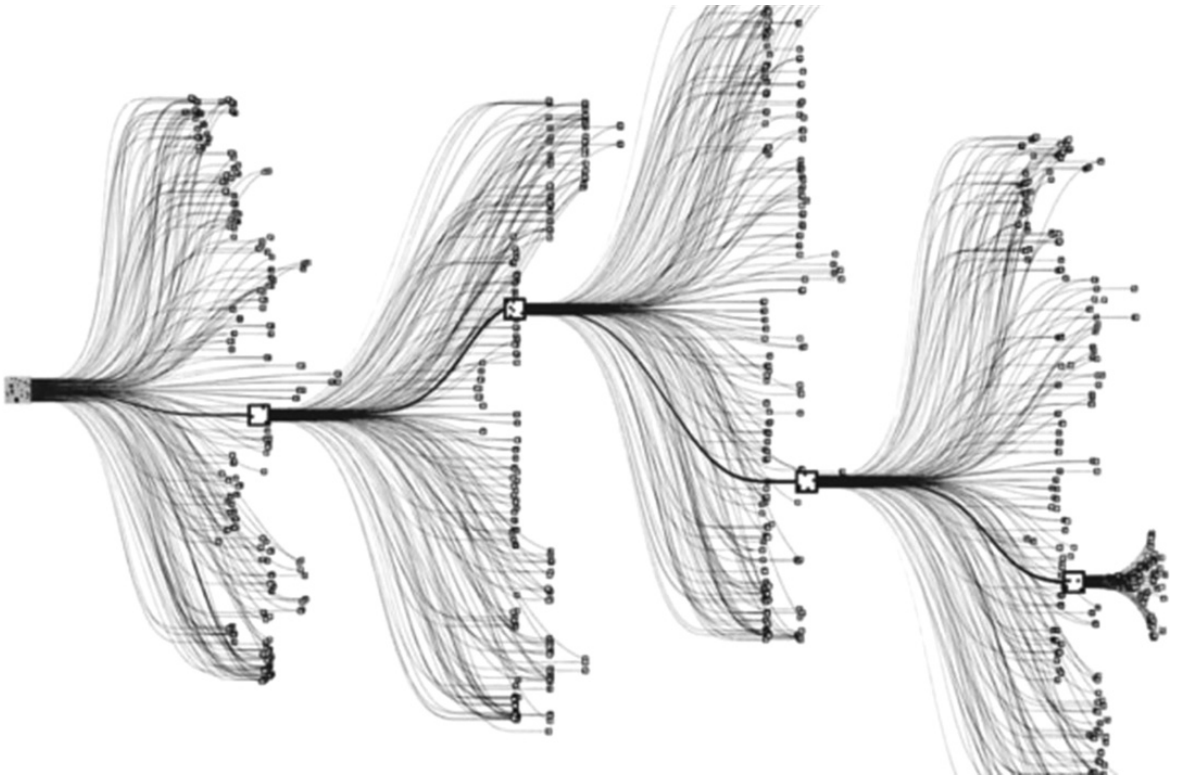

Figure 1.4 Small section of the combinatorial game tree in the game of Go, analyzed by Google's AlphaGo Zero.[12]

the goals given by its designers. Using the broad array of actuators and manipulators of its robot body, it will be able to modify its own movement algorithms and hardware, constantly optimizing its interactions with other physical objects. In short, it will have been given all the resources it needs to improve its "body" and "mind" over time – and self-improvement will be one of the prime directives inscribed into its motivational programming. What's to prevent it from launching a repeating cycle of escalating self-improvement, using its ever-growing powers to reconfigure itself over and over until it becomes unrecognizably different from its earlier designs? Some prominent AI researchers worry that humans could completely lose control over such self-evolving machines.

Equally worrisome is the prospect of AI as a military technology. In a world that's increasingly reliant on smart, automated systems – from the stock market to air traffic to the electrical grid – advanced AI will offer military officials the tantalizing gambits of cyberwarfare: a war that's over in a few minutes' time, as one nation's superior AI systems gain control over a rival's military and civilian infrastructure. This possibility – teams of AI machines and humans working together to wreak havoc – threatens to radically destabilize the military balance of the coming decades.[13]

Fossil fuels, nukes, synthetic biology, and advanced AI: the worrisome aspect, in all four cases, lies in the possibility of these powerful technologies acquiring a life of their own, yielding unforeseen or undesirable results. Each of these four inventions offers Promethean powers – with the potential for a Promethean cautionary tale wrapped inside.

IS THE DANGER REALLY THAT BAD?

To speak of existential risk in regard to any human technology might sound like hyperbole, but in fact it is not.[14] We can distinguish five levels of harm that might result from deliberate misuse or accidental malfunction in such devices, ranging from the "merely" disastrous to the genuinely apocalyptic.

1. Local disaster. Destruction of a city and its surrounding territory.
2. Cataclysm. Destruction of large swaths of a continent.
3. Return to Stone Age. Severe degradation of all human life on Earth.
4. Human extinction. Destruction of all human life on the planet.
5. Ecocide. Destruction of all biological life on the planet.

Synthetic biology and AI would probably be limited to damage at levels 1 through 4, but nuclear war or severe climate change could render the entire planet uninhabitable by most life forms for a long time.

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

The core challenge, with all these technologies, is how to reap the benefits they offer while keeping them firmly under control. Given how potentially dangerous they are, it's not enough simply to "minimize the risk" of their running amok. Even a small chance of their eluding our control is hard to justify, given that the outcome could well be the annihilation of everything we hold dear. Suppose for example that you were offered a shady business deal in which you gain a million dollars if you win, but you and your family will all be killed if you lose. Everything depends, of course, on your assessment of the underlying odds. If your chance of losing is one in a trillion, you might be willing to roll the dice for this deal. But if the odds of losing are closer to one in 20, most people would probably say "No thanks." The benefit is certainly tantalizing, but there's a real chance that things will go the wrong way and bring on a familial apocalypse. Here the risk–benefit ratio has moved closer to Russian roulette territory – a place where rational persons refuse to go.[15]

No one can know with any precision, of course, what the odds are for a catastrophe to result from one of these four high-impact technologies over the coming century. The experts in these fields vary widely in their assessments, ranging from "Relax, we've got this covered" to "Tornado coming, batten down the hatches." But if the odds appear closer to one-in-a-hundred rather than one-in-a-trillion, common sense dictates that humankind should proceed with a special kind of caution in developing such inventions.

For climate change and nukes, of course, it's already too late: the genie is out of the bottle. The overwhelming majority of climate scientists today concur in their assessment of the Earth's situation: greenhouse gases released by humans since the Industrial Revolution are accumulating in the atmosphere, causing a steady rise in average global temperatures.[16] As the planet warms, one degree at a time over the coming decades, the resultant climate change will increasingly disrupt our lives. Estimates of the threat range across a spectrum: on the "mild" end, we can expect drought, famines, flooded coastal cities, and forced migrations, while at the other end the experts warn of vast portions of the planet becoming unfit for human life. Worse still, once we cross certain thresholds or tipping points, the planet's climate dynamics could enter a self-reinforcing cycle in which the warming accelerates on its own, and the window for human preventive action will close. The problem is urgent, but humankind thus far has been tragically feckless in its response.[17]

As for the nukes, we've managed to live with them for eight decades so far without triggering a holocaust. Partly this stems from the nature of these

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

technologies themselves: they are made from rare materials, require large industrial facilities for enrichment to weapons-grade levels, and leave a telltale radioactive signature that renders them relatively easy to locate. The missiles on which they ride require a human being's decision before they are launched. Most importantly, they have been monopolized by nation-states whose behavior is subject to the rational calculus of mutually assured destruction. Taken together, these factors have allowed humankind to keep these machines from being unleashed. So far. But it's misleading to use the word "control" in characterizing our relationship with nukes. When one learns how hair-raisingly close to large-scale missile launch the world has come on several well-documented occasions – whether through machine malfunction, human error, or political miscalculation – one faces a sobering reality. Any species that voluntarily submits itself to technologies of mutual suicide, decade after decade, is playing a game that defies the logic of prudence or even sanity. Yet we just keep rolling those dice.

Synthetic biology and multipurpose AI are worrisome for a quite different reason. Once the scientific challenges of creating them have been surmounted, they will differ from nukes in one key regard: they won't be nearly as hard to manufacture, conceal, or deploy. Their development will most likely be carried out in a large number of laboratories and factories throughout the world, overseen by actors ranging from national governments to university officials to corporate executives to independent scientists. Access to these "democratized" technologies will therefore be much more widespread and volatile than has been the case with nukes, and this will render the problem of controlling them far harder.

In past centuries, the manipulation of deadly microorganisms could only be undertaken by a small number of highly trained, well-funded, and carefully screened individuals worldwide. Today we are rushing headlong into a very different world in which large numbers of persons throughout the planet will have access to the powerful tools of synthetic biology. The Harvard biologist George Church, one of the founders of the field, is bluntly honest in appraising the challenge facing us:

> For all the benefits it promises, synthetic biology is potentially more dangerous than chemical or nuclear weaponry, since organisms can self-replicate, spread rapidly throughout the world, and mutate and evolve on their own. But as challenging as it might be to make synthetic biology research safe and secure within an institutional framework such as a university, industrial, or government lab, matters take a turn for the worse with the prospect of "biohackers," lone agents or groups of untrained amateurs, working clandestinely, or even openly, with

> biological systems that have been intentionally made easy to engineer. The problem with making biological engineering techniques easy to use is that it also makes them easy to abuse.[18]

As the COVID-19 pandemic has shown, even a relatively "mild" microbial pathogen – with lethality levels closer to the flu than to Ebola – can spread globally and kill millions of persons, severely disrupting the world's economy. Many of the leading practitioners of synthetic biology believe that our society urgently needs to put in place a better system of safeguards than those that presently exist. Such safeguards would not only decrease the risks associated with synthetic microorganisms; they would also help to mitigate the ever-present danger of naturally occurring pandemics.[19]

AI poses no existential danger today: the worst thing it will probably do over the coming decade or two is take away jobs from broadening swaths of the labor market, from truck drivers to data analysts, from accountants to office workers. The economic and societal upheaval will be considerable, but the disruption stands a fair chance of remaining manageable as long as massive retraining programs and government-funded safety nets are put in place. What's more concerning is the middle-term trajectory of this technology, three to five decades out, when general-purpose machines become capable of emulating a broader range of human physical and mental faculties. At that point, the control problem becomes far more daunting, because we will be dealing with machines that exhibit sophisticated forms of human-like agency: decision-making, strategizing, forecasting, learning, adapting, networking with each other, competing and cooperating with each other, modifying themselves and their environment in increasingly powerful ways.

It's hard to overstate the significance of this development. Throughout past millennia, humans have faced two basic kinds of control problems: controlling the objects and animals that surrounded us in the material world, and controlling the behavior of other people. Of these two, it was the challenge of other people that always proved far harder, precisely because people are *agents*: they talk back, deceive you, go their own way, surprise you, outsmart you, and change unpredictably on their own. Our species became quite adept at managing the behaviors of animals and material things, and has risen to dominance on the planet as a result. Where we have always struggled, and still struggle today, is in controlling ourselves and other members of our own species.

Now we are poised (a few decades hence) to endow an entire class of machines with a broad range of capabilities specifically designed to emulate our own intelligent agency. Some perceptive observers have taken notice. In

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

2014 the Nobel-prizewinning physicist Stephen Hawking joined with Stuart Russell, one of the world's leading experts on AI, in issuing the following statement:

> It's tempting to dismiss the notion of highly intelligent machines as mere science fiction. But this would be a mistake, and potentially our worst mistake in history. In the medium term … AI may transform our economy to bring both great wealth and great dislocation. Looking further ahead, there are no fundamental limits to what can be achieved. … One can imagine such technology outsmarting financial markets, out-inventing human researchers, out-manipulating human leaders, and developing weapons we cannot even understand. *Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all.* … Although we are facing potentially the best or worst thing to happen to humanity in history, little serious research is devoted to these issues.[20]

The big challenge with AI will be to develop this technology in ways that remain reliably aligned with human interests and human values. And this will be well-nigh impossible unless we make fundamental changes in the competitive global environment – political, economic, and military – within which AI research is inevitably embedded.

WHY THESE FOUR?

Why focus on just these four mega-dangers, and not on other existential threats facing our species? Among the candidates for such a discussion, one thinks for example of a large asteroid impact, along the lines of the 20-mile chunk that hit the Yucatan Peninsula some 66 million years ago, causing a climate disruption that killed about 75 percent of plants and animals on Earth.[21] This is certainly a non-negligible category of mega-threat, and humans have put in place observatories to track such celestial objects, wagering that we might perhaps find ways to divert them from striking our planet if we can detect them early enough.[22] But these kinds of major impacts tend to happen on a million-year timescale, and it remains far from clear that we would be able to do anything about them even if we saw one of them heading our way.

The same goes for super-volcanoes. These momentous eruptions have punctuated the Earth's history from time to time, spewing masses of particulates into the skies and depositing deep layers of ash on the continents around them. The Yellowstone super-volcano erupted two million years ago, leaving a vast crater-like depression in the planet's crust; when the even larger Toba volcano in Indonesia exploded 74,000 years ago, the entire

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

planet's average temperature is thought to have dropped several degrees for a few years as a result of the "volcanic winter" that ensued. But these super-eruptions are notoriously hard to predict, and – once again – we have no clue how to stop them even if we suspect one is about to go off. The wisest policy is to keep trying to learn more about them, and perhaps to set aside stockpiles of emergency food supplies to help humankind weather the after-effects if one blows its top again.[23]

Another type of candidate is exemplified by the World War II story that opens my narrative – a powerful device that does new and potentially risky things with subatomic particles. Similar concerns recently animated a group of physicists who worried that Europe's Large Hadron Collider, situated on the border between France and Switzerland, might accidentally create a small black hole when it started running experiments in 2010 – a black hole that would instantly swallow the entire planet. As was the case during the Manhattan Project, the scientists in charge of the European collider assigned several successive teams of independent physicists to assess the plausibility of such scenarios, and their reports convincingly ruled out any doomsday outcome from the giant machine's operations.[24]

During the early 2000s, the nascent field of nanotechnology also elicited alarums among some scientists, who feared that (hypothetical) molecular nanomachines might someday unleash themselves across the continents, multiplying exponentially and converting all the planet's matter into copies of themselves. This came to be known as the Gray Goo Scenario – for in that case all that would be left of Earth would be a colorless blob of satiated nanomachines. This hypothesis, too, was thoroughly assessed by a variety of experts, and ultimately dismissed as either vanishingly improbable or downright impossible.[25] (I'll return to nanotechnology later on when I discuss self-replicating bioengineered microbes.)

The four technologies discussed in this book fall into a quite different category. Each of them could plausibly kill a great many of us within the next hundred years (or ten minutes from now, in the case of nukes) – and we are definitely in a position to decrease the risk of their running amok. Although experts disagree about the exact nature of the dangers they pose, there is clear consensus that we need to take them seriously, and find better ways of governing their development and deployment.

WHAT CAN A HISTORIAN CONTRIBUTE TO THE DISCUSSION?

The inventor and futurist writer Ray Kurzweil (who also works a day job as a Director of Engineering at Google, Inc.) firmly believes that the

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

exponential forms of technological innovation I've been describing are coming our way whether we like it or not. Under the heading "The Inevitability of a Transformed Future," he writes:

> The window of malicious opportunity for bioengineered viruses, existential or otherwise, will close in the 2020s when we have fully effective antiviral technologies based on nanobots. However, because nanotechnology will be thousands of times stronger, faster, and more intelligent than biological entities, self-replicating nanobots will present a greater risk and yet another existential risk. The window for malevolent nanobots will ultimately be closed by strong artificial intelligence, but, not surprisingly, "unfriendly" AI will itself present an even more compelling existential risk.[26]

Humankind, in this picture, has landed itself on a sort of historical treadmill of ever-rising technological powers, each of which presents a more daunting level of intertwined benefits and dangers. Last year's beneficial breakthrough becomes today's deadly threat, requiring the urgent invention of today's new beneficial breakthrough, which in turn will become tomorrow's deadly threat – and so on escalating without end. "The meta-lesson here," Kurzweil acknowledges, "is that we will need to place twenty-first-century society's highest priority on the continuing advance of defensive technologies, keeping them one or more steps ahead of the destructive technologies."[27] Apparently unfazed by the underlying ironies of this self-imposed rat race, Kurzweil is in fact echoing the sentiments here of prominent critics of technological modernity such as Lewis Mumford, Jacques Ellul, Ivan Illich, Wendell Berry, or Bill McKibben.[28] But instead of adopting their critical stance, he shrugs his shoulders and says, in effect: "Let 'er rip! We must innovate ever faster or die."

Yet this is a fallacy. In reality, there is nothing inevitable or predetermined about the historical process that generates these technologies, any more than the development of the bicycle or the internal combustion engine was predetermined. New technologies do not just suddenly appear in our midst, arising fully formed out of the labs of scientists and inventors, forcing us humans to adapt to the novel capabilities with which they endow us. Historians of science and technology have persuasively shown that this is a misguided way to view innovation; they refer to it as the fallacy of "technological determinism." Here is the way the historian David Nye sums it up in his book *Technology Matters*:

> A technology is not merely a system of machines with certain functions; rather, it is an expression of a social world. Electricity, the telephone, radio, television, the computer, and the Internet are not implacable forces moving

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

> through history, but social processes that vary from one time period to another and from one culture to another. These technologies were not "things" that came from outside society and had an "impact"; rather, each was an internal development shaped by its social context. No technology exists in isolation. Each is an open-ended set of problems and possibilities. Each technology is an extension of human lives: someone makes it, someone owns it, some oppose it, many use it, and all interpret it.[29]

Consider for example the history of airplanes.[30] The first powered flight in a heavier-than-air machine lasted twelve seconds and carried Orville Wright 120 feet over the sands of Kitty Hawk. The craft's velocity: 6.8 miles per hour. Over the half-century that followed, as aircraft designs went through a spectacular series of transformations, their performance grew dramatically. By the end of the Second World War – a mere four decades after Kitty Hawk – the jet-powered Messerschmitt 262 was hitting speeds above 540 m.p.h., and the B-29 bomber boasted a range of 5,500 miles. Extrapolating from this giddy upward trajectory, some sci-fi writers in the 1950s confidently predicted that by the year 2000, sleek hypersonic craft would be whisking passengers from one continent to another with the ease of a cross-city bus ride. Yet it was not to be. Even though exciting supersonic planes like the Concorde were indeed built in the 1970s, they proved far too expensive for airlines to operate. International air travel did grow impressively during the postwar decades, but the economics of the industry, along with rising environmental concerns, took aircraft designs in a quite different direction: the new norm was cheap seats on large planes powered by efficient engines. Even though supersonic flight has definitely lain within the technical reach of aircraft engineers, today's most advanced passenger aircraft like the Airbus A380 and Boeing 777 are designed primarily to save gas, emit fewer pollutants, make less noise, and safely (if boringly) carry large numbers of passengers to their myriad destinations. They fly no faster than the airliners of the 1960s.

This is a crucial point for my argument in this book. Just because powerful and exciting new technologies are scientifically feasible, this does not mean that their creation is unavoidable. Humans make choices, both as individual consumers and as societal collectivities, and these aggregated choices exert a strong shaping effect on the development of our inventions. Some of our devices succeed and proliferate, becoming a part of our everyday lives; some assume strange and unexpected forms; some get used for quite different purposes than the ones for which they were designed; and some never get built at all. This shaping process is rarely simple and straightforward. It results from millions of choices made at many

levels of a complex societal system, and its unfolding is hard to predict. But at bottom, we humans do have a significant say in how our technologies develop – what kinds of machines we create, how rapidly we assimilate them, and how we choose to use or not use them.

It's here that nontechnical writers – historians, economists, political scientists, philosophers, and others – have something helpful to offer. Precisely because all new technologies come into being within a broader societal context, humanists and social scientists can offer a complementary dimension of analysis that renders the technical solutions more effective. Consider for example nuclear weaponry. One side of the story is primarily technical in nature: the development of hydrogen bombs, tactical nukes, submarine-based missiles, cruise missiles, command-and-control systems, and so on. The other side of the story, equal in importance, is the Cold War historical context within which these weapons were funded, developed, and deployed. Here we need to take into account the logic of arms races, the fluctuating relationship between the superpowers, the internal politics within nuclear nations, the military-industrial complex, nuclear proliferation, arms control efforts, and the ideological systems that were used to justify the deployment of such horrific weapons.

Even the most creative attempt to reduce the nuclear danger would fail if it focused solely on technical solutions: it could only succeed if it also took into account the societal forces that shape the nuclear equation. For example, US President Ronald Reagan came up with a brilliant technological fix with his Star Wars initiative in the early 1980s – a shift from "mutually assured destruction" to "mutually assured defense." Instead of racing to find new ways to blow each other to smithereens, the superpowers would focus their ingenuity on building impenetrable shields to protect their territories from nuclear attack. Yet Reagan's path-breaking idea went nowhere, precisely because it failed to embed his proposal within a new international system operating beyond the strategic rivalry of the Cold War. If the radical technical solution was truly going to work, it would have needed to go hand-in-hand with an equally radical *political* shift in how countries pursued their security. My approach in this book is to continually bring these two dimensions of analysis together, embedding the most promising technical solutions within the political and historical frameworks through which they will have to be implemented.

THE BOOK'S ARGUMENT IN A NUTSHELL

1. Four major catastrophic risks confront humankind over the coming century. The cumulative danger is higher than ever before, because climate

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

change and weaponized AI now pose serious new threats in addition to the older threats of nuclear war and pandemics.

2. Past efforts to mitigate these kinds of mega-dangers have included modest initiatives such as climate treaties, arms control deals, or limited pandemic precautions, as well as bolder moves like the US government's 1946 proposal for international control of atomic weapons, the 1972 Biological Weapons Convention, or Ronald Reagan's 1983 missile shield initiative. While these were important steps in the right direction, they have fallen far short of what is needed. In all cases, the most salient problem has been the way every nation continues to fend for itself in a ruthlessly competitive world arena.

3. An effective response to these four mega-dangers will require moving beyond the international self-help system and creating coordinated instruments of global governance. These instruments can be progressively strengthened in successive phases of institutional change over the next century and a half. The historical backdrop for this transformative process lies in the rapid growth of international laws, supranational institutions, and globe-spanning networks over the past 120 years – from the League of Nations in 1920 to the United Nations in 1945 to today's International Criminal Court and European Union.

4. Moving toward a more tightly integrated global framework strikes many people as a utopian pipe-dream. Yet creating such a framework would not require a radical departure from established practices, but only a continuation and extension of the remarkable cross-border initiatives and institutional innovations that humankind has already been undertaking over the past century.

Most people, when they think of world government, conjure two kinds of images: a tyrannical superstate along the lines of Huxley's *Brave New World*, or the feckless, anemic United Nations of today. The emotional tone is either fear or disdain. But it doesn't have to be that way. A world government could be built according to a very different blueprint – one that enhances our freedom of action rather than constrains it; empowers its citizens and opens new doorways for flexible connection and teamwork; offers effective instruments for resolving the deep conflicts that inevitably arise among citizens and groups; remains transparent and accountable at all levels of its functioning; establishes a state of law at the global level – freely chosen laws that the world's peoples have negotiated amongst themselves; and imposes nothing from above except the minimal necessary instruments

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

for enforcing those laws and keeping the peace. What if a world government offered vibrant new forms of self-rule, political expression, and liberation to humankind?

This may all sound too good to be true, but in fact many of the instruments for such a global framework are already beginning to emerge today. We tend to apply the clunky, top-down metaphors of a 1950s bureaucracy when we think of world government, but what if instead we apply the bottom-up metaphors of the most responsive, decentralized systems of the present day? What if, instead of thinking *Encyclopaedia Britannica*, we think Wikipedia? Our deep challenge, over the coming century, lies in learning how to take up these new kinds of organizational instruments and apply them creatively across the oceans and continents.

5. Planet-level solutions for today's world should focus on reducing the most grievous risks, while developing new forms of cross-border concertation and regulation. Examples include: cooperative pacts among select groups of nations for swift decarbonization of energy systems; treaties that restrict new and destabilizing weaponry; boosting global pandemic preparedness; government oversight for synthetic biology; promoting safety research in AI.

6. Chapters 15 through 20 describe a series of incremental steps toward a federal framework for global democratic governance. These concrete measures could be undertaken over the span of many decades, gradually putting in place a new set of planet-level institutions and norms. The key challenges here are: creating a political system that allows dictatorships and democracies to work constructively together; revamping the UN Security Council so it more accurately reflects the realities of global economic and military power; developing an equitable system of weighted voting in UN institutions, so that key policies can be implemented effectively; reducing the gross disparities in wealth and opportunity that divide the world's peoples; keeping the UN system rigorously accountable and transparent in its operations; and building robust instruments of collective military security and economic sanctions, capable of dealing decisively with rogues, cheaters, or fanatics. In this federal structure, existing national governments would continue to do most of the day-to-day running of people's affairs; only the truly global matters such as military security, climate change, or regulating dangerous technologies would be assigned for coordination by the UN and its affiliated bodies.

7. Fortunately, this transformative process will not be an all-or-nothing proposition. The benefits of international cooperation grow by degrees, in

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

direct proportion to the cross-border solutions that have been put into practice. This means that even relatively small cooperative innovations can start making an impact right away in delivering heightened stability, prosperity, and security; and as these basic measures are followed up with more intensive forms of institution-building, the benefits increase progressively in scope. Humankind does not have to build a full-fledged system of global federal governance in order to start working on these problems successfully: we can start making a tangible difference right now.

8. Generating the political will for this incremental restructuring of world politics poses a defining challenge for our time. We can draw inspiration from the partial success of two major recent initiatives: the green movement and the creation of the European Union. Like the international activists who launched these two endeavors in the 1950s, our generation can turn to a wide array of powerful tools for mobilizing change at all levels of society, from citizens to leaders, from the local to the global, from institutions to habits of mind. Both of these remarkable stories can offer valuable lessons as we press forward to make the necessary changes in world politics.

9. The timeframes listed in the table of contents – culminating in a mature governance system by 2150 – are of course speculative in nature. I chose 2150 as a benchmark because it conveys the likelihood that this will be a long, slow process, requiring the gradual accumulation of myriad smaller changes in world affairs. Strong incentives – both sticks and carrots – are likely to motivate humankind as this process unfolds. Penalties for partial failures could include a hotter planet, nuclear conflicts or accidents, bioengineered pandemics, or AI disasters. These are some of the harsh "lessons" we hopefully won't have to learn along the way. The benefits of partial successes, on the other hand, would be tremendous: a cooling planet; an end to arms races in weapons of mass destruction; a phased reduction of standing armies, yielding a massive peace dividend that could be used to reduce world poverty; and a robust control system for advanced bioengineering and AI. These kinds of penalties and benefits could impel major innovations and reforms in a similar way to the impacts of World War I and World War II – and their cumulative pressure will only escalate as the decades go by.

10. The path-breaking achievements of the past century offer grounds for cautious hope. Over the coming decades, economic and technological interdependence will continue to intensify, binding the interests of the world's peoples even more tightly together than today. As nations and

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press

regions find themselves increasingly "in the same boat," win–win solutions among them are likely to become more self-evidently attractive.

To be sure, the possibility of failure is ever-present and very real. Building more effective instruments for planet-level governance will prove exceedingly hard, and we can expect plenty of disheartening setbacks along the way. But the evidence from the past, and the plausible trajectory of challenges and incentives that await us in the coming century, suggest that it can be done.

* * *

In order to lend more concreteness to the discussion, I offer eleven fictional vignettes in which I envision what the future world of mega-dangerous technologies might actually look and feel like – and how they may be brought under control. As long as we bear in mind that these vignettes are speculative exercises – educated guesses – they can help illustrate the kinds of challenges that citizens of the coming century may find themselves encountering.

https://doi.org/10.1017/9781009160315.001 Published online by Cambridge University Press