Wisdom at the Brink

A planet-level strategy
for dangerous technologies:

nuclear, nanotech, synthetic biology, AI

DRAFT CHAPTER: PLEASE DO NOT CITE OR SHARE WITHOUT THE AUTHOR'S PERMISSION

It might be a familiar progression, transpiring on many worlds.

> A planet, newly formed, placidly revolves around its star;
> life slowly forms;
> a kaleidoscopic procession of creatures evolves;
> intelligence emerges…
> and then technology is invented…
> Science, they recognize, grants immense powers.
> In a flash, they create world-altering contrivances.

Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others, not so lucky or so prudent, perish.

– Carl Sagan, *Pale Blue Dot* (1994)[1]

Contents

Chapter Nineteen

Plan B for strong AI:

coexisting with superintelligent machines

We know from hard experience what a city looks like after it's been nuked. We know from the 1919 influenza what it means to have fifty million people die in a viral pandemic. But we have no clue what will happen when the first superintelligent AI is built. All we can do is listen to the experts as they make educated guesses, and piece together their hopes and fears into a rough sketch of what may possibly lie ahead. Their visions can be broken down into six nested questions:

- Is it possible to build a superintelligent machine?
- If it proves possible, will humans build one?
- If humans build one, with the takeoff be hard or soft?
- If the takeoff is soft, will a global hegemon emerge?
- If a hegemon emerges, will it be malevolent or benevolent?
- If no hegemon emerges, how will humans relate to these new machine-beings?

Is it possible to build a superintelligent machine?

The simplest answer is that we don't know for sure. Some researchers in the field believe that humans will never be able to create such a machine, but most are confident that superintelligence lies within the plausible reach of our scientific and technological skills, and will become feasible sooner or later. The range of estimates goes from a few decades to a couple centuries in the future, with a significant number predicting that such machines could well be with us by the year 2075.[2]

If it proves possible, will humans build one?

I argued in Chapter Four that humankind should refrain from building such machines unless they can be shown to be reliably stable and obedient. On the other hand, I also acknowledged that strong incentives will exist to shove aside all restraint: these machines will confer immense power on their creators, and the logic of arms races will be hard to resist. It would certainly be a welcome achievement if humankind were to show the collective wisdom required to keep postponing the creation of such machines

indefinitely, until their safety and security can be guaranteed. But the historical track record suggests that we shouldn't bank on it. It's also unclear whether it will ever be possible to guarantee in advance that such machines can be designed for reliable stability and obedience.

In the absence of effective regulation – i.e., if we keep going the way we're going – what seems likely to happen is a gradual, messy series of innovations and incremental improvements in machine capabilities, progressively blurring the lines between narrow AI and AGI (artificial general intelligence). Once machines begin crossing the threshold of AGI, many experts believe such powerful, versatile machines will be able to launch the recursive, accelerating cycles of self-improvement that lead to exponential gains in physical and mental powers – whether in cooperation with humans or on their own initiative.[3] A breakthrough into machine superintelligence is the logical culmination of such a process.

If humans build one, will the takeoff be hard or soft?

The transition from AGI to ASI (artificial superintelligence) could plausibly happen along two broadly different pathways. One pathway would be swift and radical, causing a major rupture in the fabric of history: this is known in the expert literature as a hard takeoff or intelligence explosion. The other pathway would be much slower and more incremental, and would result in the genesis of relatively stable machines that do not undergo (or undertake) exponential cycles of self-improvement. This is known as a soft takeoff.

I've depicted one possible scenario for a hard takeoff in the vignette in Chapter Two about *Tóngzhì*, the android AGI that works with its human creators to modify itself through successive cycles of redesign into a state of nascent superintelligence. In that vignette, I deliberately avoided the standard sci-fi imagery of runaway AI as a malicious force, and portrayed *Tóngzhì* as having a neutral impact on human civilization – a kind of demigod whose nature inclined it toward peaceful coexistence with the rest of the planet. *Tóngzhì* rapidly outgrows its android body, transforming itself into an intelligent, powerful, ubiquitous cloud of evolving agency that spreads into the material world and appears set to merge with the biosphere. But of course, this is just one possible plot line for the story. *Tóngzhì* could turn out malicious, or actively benevolent, or (perhaps more likely) could ignore us biological creatures in the same way that you and I ignore the bacteria in our gut. It could refashion the planet in ways that allow us to continue to exist, or in ways that end biological life and pave the way for a new Postbiocene Epoch.

At bottom, therefore, the four most important qualities of a hard takeoff – at least, from a human perspective – are the following:

▪ It is possible – and perhaps even probable – that a hard takeoff would result in the creation of an exponentially self-modifying superintelligent machine.

▪ Such a machine could rapidly acquire immense capabilities for radically altering the material world.

▪ We humans would probably be completely at the mercy of such a machine, finding ourselves incapable of preventing it from doing whatever it does.

▪ We have no way of knowing in advance what such an entity would end up doing to us, to the world, or even to itself.

Creating such a machine amounts to rolling dice with the fate of the biosphere, in total blindness. It's even possible that the consequences of such an act of creation would radiate outward from our planet, affecting the rest of the solar system and perhaps the galaxy. To call such an act of creation "immoral" is to strain the capabilities of human language.

In the foregoing chapters I've sketched some elements of the regulatory system that will be needed to ensure that a hard takeoff does not take place. As we saw, the challenges involved in creating such a protective system will be wickedly difficult ones, particularly at the international level, where new forms of global cooperation will have to be put in place. But the bottom line is clear: preventing a hard takeoff should constitute one of the core goals of humankind over the coming century – on a par with avoiding pandemics, global warming, or nuclear war.

The other plausible pathway – a soft takeoff – might perhaps take shape along the lines of the benevolent Synthia robots I described in Chapter Four. In this scenario we are assuming that researchers find a way someday to build AGI machines that do not undergo an intelligence explosion, but remain relatively stable and obedient over time. I have no clue how the AI designers might actually go about achieving this result, for it would require resolving the inherent tension between autonomy and obedience that I described in Chapter Six. Nevertheless, it is conceivable – at least in principle – that radically new motivational architectures and other design breakthroughs might one day render such machines feasible, so it is worth taking a moment to envision what that world might look like.

These Synthia robots would be most likely to mesh well with human society if they were designed as androids (more or less), but they would undoubtedly work in concert with myriad other forms of stationary AI machines, as well as drones, wheeled bots, swarmers, crawlers, swimmers, tunnelers, fine-grained manipulators, nanobots, and other specialized gizmos. The Synthias would ideally be equipped with rich commonsense knowledge of our human life-world, as well as a motivational structure that inclined them to be unfailingly friendly toward humans and eager to please. There would probably be thousands, if not millions, of such intelligent and versatile robots operating among us, and they would no doubt be linked in communication networks that allow them to coordinate their activities so as to serve us more effectively. They would be superintelligent, in the sense that their cognitive capacities would outstrip those of humans in many domains, but they would still remain reliably subservient to humans when it came to determining their basic goals and purposes. They would be semiautonomous, but their functioning would always remain subject to countermanding decisions by authorized human operators. They would not use their impressive ingenuity to improve or redesign their own hardware or software, except within narrowly-defined parameters overseen by humans.[4]

Which pathway is more likely – hard or soft takeoff? The answer depends on a number of variables. If humans succeed over the coming decades in getting beyond the self-help system of international politics, shifting gradually into a more cooperative mode of global governance, then the prevention of a hard takeoff becomes proportionately more feasible. But if we persist indefinitely with roughly the same forms of international competition and weak technological governance that prevail today, then sooner or later a hard takeoff becomes a much likelier outcome.

The soft takeoff scenario, for its part, depends on how our society allocates resources to AI research over the coming century. If we continue to direct most of our efforts toward building ever more powerful and multifunctional machines, then we are likely to witness a hard takeoff before AI designers find a way to render our machines reliably stable and obedient. If, on the contrary, we start devoting major funding and talent toward AI safety and security – and give these features a higher priority than power and flexibility – then the plausibility of a breakthrough into a world of benevolent AI machines goes up. It's up to us to decide, through our individual and collective choices, which pathway our future will follow.

If the takeoff is soft, will a global hegemon emerge?

Whoever makes the final breakthrough into building stable and obedient AGI machines stands a good chance of swiftly becoming one of the most powerful actors on the world stage. Whether the successful inventor is a single research team, a corporate lab, or a governmental task force, the persons who control such machines will acquire immense leverage over the physical, economic, and social world. Such machines will offer remarkable new capabilities in all the following domains:

- Devising new strategies for making economic gains
- Manipulating the physical world, either directly or indirectly through other machines
- Controlling basic infrastructure such as electric grids, water supplies, air traffic, etc.
- Control over cyberspace
- Designing new machines
- Offering powerful new methods for persuading or influencing other people
- Achieving communications and transportation breakthroughs
- Making scientific discoveries
- Solving wicked problems like global warming, disease, or poverty

These kinds of factors are precisely the ones that influence the global ranking of geopolitical actors today, marking the difference between superpowers, great powers, and lesser powers. Whoever controls the networks of AGI machines will tap into a new source of leverage that outweighs all other factors, precisely because it will hold the key to so many of them.[5]

A pivotal question, therefore, has to do with the way such powerful machines are introduced into society. If they come into being as the exclusive property of a single corporate or national actor, and if that actor proves willing to press this monopolistic advantage, then we will probably witness the rapid emergence of a single dominant player in world affairs. Such a technological hegemon could easily use its newfound

dominance to actively suppress all rivals who were close to building a competing AGI network of their own. As the years went by, it is likely that this hegemon would be able to further consolidate its power, nipping all challenges in the bud, and establishing an unassailable position of economic, military, and cultural supremacy.[6] In such a world, even nuclear weapons would probably cease to offer the kind of absolute guarantee of independence that they do today, because the warheads and missiles themselves would lie within easy reach of the dominant network of AGI machines – either directly (through coerced control) or indirectly (through AGI-facilitated hacking of the weapons' command-and-control systems).

But things may also go quite differently. It's possible that the advent of AGI technology will happen gradually, through a messy, uncoordinated series of incremental breakthroughs interspersed here and there among many nations, corporations, and research teams around the world. This broad distribution of key innovations would undermine the ability of any single player to achieve a decisive advantage over the others. In this scenario – since each player exerts control over a particular aspect of AGI technology, and no single player controls all aspects – one would expect to see a fluid balance-of-power system emerge, with multiple actors continually jockeying for relative position. If history is any guide, such systems are notoriously unstable, and can yield a wide variety of outcomes. One outcome would be the gradual emergence of a single dominant player; another would be the banding together of many players to prevent a single player from predominating; still another would take shape as an indefinite continuation of balance-of-power dynamics, with multiple AGI systems emerging simultaneously in different parts of the world, each roughly equivalent in prowess to the others.[7]

Elon Musk considers the possible emergence of an AGI-based global hegemon a clear and present danger over the coming decades.[8] In 2015, he put his money where his mouth was, joining with several other high-tech pacesetters (including PayPal founder Peter Thiel, Indian tech giant Infosys, Microsoft, and Amazon Web Services) in creating OpenAI, an organization whose primary goal was to ensure that if AGI becomes a reality, it will be shared as equally as possible by all humankind.[9] Together, they pledged up to $1 billion for this endeavor. "Artificial general intelligence," they wrote in their mission statement,

> will be the most significant technology ever created by humans. … We believe AI should be an extension of individual human wills and, in the spirit of liberty, as broadly and evenly distributed as possible. … As a non-profit, our aim is to build value for everyone rather than shareholders. Researchers [at OpenAI] will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world.[10]

Musk and his colleagues acknowledged the paradox at the heart of their project: they were working furiously to build the world's first AGI, so as to share it with the rest of humankind and thereby prevent a single hegemon from emerging; but this frantic effort would probably hasten the arrival of AGI, and thereby inadvertently increase the danger of a hard takeoff.

> We are concerned about late-stage AGI development becoming a competitive race without time for adequate safety precautions. Therefore, if a value-aligned, safety-conscious project comes close to building AGI before we do, we commit to stop competing with and start assisting this project.[11]

In other words: *we care more about the fundamental safety of AI and the long-term welfare of humankind than we do about being the winners in this race.* For AI researchers, taking this ethical stand might be compared with physicians adopting the Hippocratic Oath, or genetics researchers adopting the Asilomar principles: OpenAI was leading by example, seeking to set the moral ground rules for the AI field.[12] It was a bold and constructive step toward the kind of cooperative ethos that will need to prevail, if humankind is to stand a realistic chance of controlling the technologies of apocalypse.

Still, it will always remain possible for a less scrupulous team of AI researchers to refuse this kind of altruism and self-restraint. In the absence of effective global regulation and enforcement, such a team may well succeed in cornering the market on AGI before anyone else can intervene. Precisely because such an outcome cannot be ruled out, it behooves us to consider the emergence of a global hegemon as a serious possibility.

If a hegemon emerges, will it be malevolent or benevolent?

Everything will hinge on the character, ideology, and cultural background of the persons wielding the AGI monopoly. Decades of research into the psychological and ideological makeup of Germany's population in the Nazi era, coupled with exhaustive studies of persons following other extremist ideologies such as fascism, communism, or religious authoritarianism, all point in the same general direction: people can talk themselves into doing all kinds of vicious things if they frame it through the logic of "the ends justify the means."[13] Hitler and Stalin were murderous sociopaths, but many of their followers were not: they were ordinary citizens who, for all kinds of mundane reasons such as peer pressure, careerism, prejudice, misplaced idealism, conformism, or just plain cowardice, allowed themselves to participate actively in the bestial acts that blighted the history of their time. It turns out that one doesn't have to be a sociopath to support a political system that brutalizes millions of other humans.[14]

We don't have to strain our imaginations, therefore, in order to envision what a malevolent AGI hegemon might look like. There are plenty of historical episodes to choose from: Spain under the Inquisition, Stalin's Russia, Nazi Germany, Mao's China, Pol Pot's Cambodia. To these examples we can add the vivid sci-fi literature of dystopia: Orwell, Huxley, and others. In order to be fully accurate, however, we would need to tweak these dismal depictions a bit further, taking into account the full power of AGI harnessed to achieving the goals of a police state. The oppression of such a system might be blatant, as depicted in Orwell's *1984*, or it might be more subtle, as Huxley envisioned in *Brave New World* – a social order in which most people are perfectly content with their own dehumanization and enslavement.[15] But these two archetypal novels probably still fall short of depicting how awful a malevolent AGI hegemony could be in practice. In

Orwell's and Huxley's dystopias, some of the main characters are still capable of thinking for themselves and rebelling against the oppressive system that holds them in thrall. But in a police state controlled by AGI – a social order in which powerful AI is everywhere, and people's bodies (and brains) can be inscribed with monitoring devices (and possibly with mind-altering devices) – it is hard to see how the oppressive system can ever be brought down.[16] "Resistance is futile," said the cyborg enslavers in the famous Star Trek imagery of the Borg – but in a malevolent AGI hegemony, the more accurate phrase will probably be: "Resistance is inconceivable."[17]

But what if the persons wielding the AGI monopoly turn out to be not at all like Stalin or Hitler? What if they are a group of reasonable, democratically-minded, ethically conscious individuals whose goals are sincerely focused on promoting human welfare? Here, too, the historical track record offers striking examples to ponder. Between 1945 and 1949, the United States enjoyed a monopoly on nuclear weapons, and could conceivably have used this radical military advantage to establish a globe-spanning imperium under Washington's thumb. The philosopher Bertrand Russell publicly advocated precisely such a move by the U.S. – a preventive war against Soviet Russia, arguing that American domination of the planet was preferable to the danger of a nuclear holocaust, which was likely to happen if the Russians were allowed to get the Bomb.[18] But Harry Truman and other U.S. officials showed no inclination to forge a global empire at gunpoint: they were certainly keen to influence world affairs, but preferred to do so through more indirect means such as commerce, diplomatic pressure, and military coalitions with like-minded partners. The character, ideology, and cultural background of American leaders rendered the idea of forcible global hegemony utterly repugnant: so instead of pre-emptive nuclear war, they launched the Truman Doctrine and Marshall Plan.[19]

Another intriguing episode took place in the Soviet Union between 1985 and 1989. Here was a government that enjoyed an ironclad monopoly on domestic political power through the Communist Party and secret police, and exercised absolute military dominance over its empire of satellite countries in Eastern Europe. And yet, the leadership group around Mikhail Gorbachev came to believe that this imperium was both morally bankrupt and doomed to economic and technological decline – so they voluntarily abandoned their monopoly on power and opened up Russian politics to genuine democratic processes. At the same time, they also told the East Europeans that the U.S.S.R. would no longer impose its will by means of tanks: they were free henceforth to choose what sort of government they wanted. The character, ideology, and cultural background of this group of Russian leaders was antithetical to the brutal forms of domination inherited from the Stalinist era, and as a result, the mighty Soviet empire swiftly unraveled, with astonishingly little violence.[20]

Historical examples like these, in which powerful groups of leaders rejected the option of forcible hegemony, even when it lay within their grasp, suggest that an AGI monopoly need not necessarily devolve into a totalitarian nightmare. Everything depends on the nature of the persons who control the AGI machines. What would it mean, then, for a small group of individuals to wield preponderant power over the world, but to do so through an ethos of liberty, human rights, and democratic values?

Vignette 7: Life under the tutelage of a superintelligent hegemon (2114)

To hell with the naysayers. The invention of the benevolent Synthias is the best thing that's ever happened to humankind. All these people carping about the lack of freedom just don't get it: today we have all the freedom we can realistically afford.

I mean, just look at the benefits. Fifty years ago, no one would have dreamed we could live in a world like this. The robots are everywhere, doing our bidding, constantly thinking up new ways to make us safer, healthier, happier, more prosperous. Poverty is nearly gone, now that even the least developed countries have joined the Basic Income Network. Every citizen on the planet gets his or her own monthly allotment, more than enough to live comfortably.[21] *Education and health care are universally free: my son Bob went to college in Switzerland, my daughter Kathy went to New Zealand for her medical training. They have classmates from places like Zimbabwe, Venezuela, or Bangladesh, where rising prosperity has opened up unheard-of new opportunities for everyone. I left my job at Boeing twelve years ago and ramped up my art work from a hobby to a full-time project: next month I'll be presenting my paintings for the first time at the Seattle Art Expo!*

The three major threats to human survival have all been defeated. Global warming is now in reverse, ever since the Synthias built the solar shield at the Lagrange point between our planet and the sun. Nuclear war is impossible, now that the Freedom Party has a global monopoly on military force. And even though we've come hair-raisingly close several times, we've managed so far to avoid an intelligence explosion: the network of Synthias has caught the AI criminals each time, before they could unleash an uncontrollable ASI on the rest of us. Those bastards deserved the death penalty, in my opinion: that's one point on which I disagree with President Finwood and the Freedom Party's leadership. Rotting in prison is too good a fate for someone who plays dice with the fate of the entire planet.

We've made it successfully through the most dangerous technological and political transition in human history, and these clowns on the left-wing and right-wing fringes still aren't satisfied. "It's a global dictatorship," they say. "We didn't have any choice in setting up this system." "There are no internal checks and balances." "They just pretend to listen to us and then do whatever they want." "The Freedom Party is a clever façade, but it's the machines who are really calling the shots." "Why aren't we allowed to hold real elections?"

On and on they go with their endless criticisms, blabbing on talk radio, fundraising, organizing rallies and reform movements – does anybody stop them? Have they been arrested? Hell no. The next day comes and there they are again, carping and criticizing to their heart's content. That sure looks like freedom to me! Karen Finwood said it on TV just two nights ago: the opposition parties have every right to organize and seek reforms, and the government will listen, and it will take their recommendations seriously.

What more can you ask than that?

They call people like me naïve. They say we're dupes of a manipulative, paternalistic system. Well, if that's true, then the majority of the citizenry are dupes. The bottom line is this: with the technologies as advanced as they are, both for good and for evil, there has to be a single party at the top of the power pyramid. Somebody has to set the rules and enforce them – otherwise you get chaos, and out of that chaos you can bet your life some bunch of sociopaths will build a runaway superintelligence and destroy us all. It's as simple as that. Either you're with the Freedom Party, or you're in favor of self-destruction.

Once the AGI machines were invented, there was no choice anymore. Whoever controls those machines controls the world. We can thank God it was good people like Finwood and her crew who happened to be the ones who made the final set of breakthroughs.

Some folks say there are constant power struggles within the Freedom Party – people trying to elbow their way into Finwood's entourage, or other people plotting to wrest power away from her altogether. But those are just rumors – they've been saying things like this for the past thirty years. And there she is, one day after the next.

Thank God for President Finwood. Honestly, I do believe I'd take a bullet for her. If a nearby Synthia didn't beat me to it.

* * *

To a hundred-year-old Chinese woman living in the present day, the foregoing vignette might not sound so bad. As she takes stock of the situation in her country, she admits that the stern paternalism of the Communist Party restricts her freedoms: she sees that her children and grandchildren are not at liberty to say or write whatever they please, and that their economic options are firmly constrained by the central government. From her perspective, though, the overall improvement compared with the nightmarish turmoil of the mid-twentieth century is stark and undeniable – and for this she is deeply grateful. Today's society isn't perfect, but she knows from direct experience how much worse it could be.

In the eyes of this Chinese woman, the benevolent paternalist governance of the Freedom Party, as depicted in the vignette, might not look unfamiliar. She might find herself in agreement with our retired Boeing engineer: better to live under a safe, prosperous paternalist dictatorship than to face the dangers of a freer, more open, and therefore more turbulent system.

On the other hand, to a hundred-year-old woman from a democratic nation like Britain, France, Sweden, or the U.S., the foregoing vignette would probably evoke reactions of revulsion. Having experienced a long life under basic political and civil liberties, she would regard the forcible constraints imposed by even the most well-intentioned group of leaders as an unacceptable despotism. The lack of democratic accountability, coupled with the absence of internal checks and balances, would make the reign of the Freedom Party seem like a soft totalitarianism. She would no doubt accuse our retired Boeing engineer of having a accepted a devil's bargain.

But humankind is by no means fated to experience the rise of an AGI-induced global hegemon – whether it be malevolent or benevolent. Other pathways, less drastic in nature, may present themselves.

If no hegemon emerges, how will humans relate to these new machine-beings?

If stable, obedient AGI technology comes about slowly and incrementally, and no one succeeds in monopolizing it, then the door will be left open for a world without a single hegemonic power. We would then be faced with a fluid, multipolar geopolitical order not all that different from today's global society. The big question, in such a future, will be whether humankind stays indefinitely with the international self-help system that prevails today, or succeeds at gradually ramping up the institutions, practices, and mental habits of cooperative security that I described in Part III.

Let us envision for a moment a multipolar world of the mid-21st century, with no AGI monopoly, in which nations compete with each other as they do today. One of the main characteristics of such self-help systems is that they are inherently unstable. Everyone is constantly jockeying for position, nervously eyeing the resources and prowess of the other players. In such a system, to opt out of the race is itself a tacit choice to accept a subordinate role. Thomas Hobbes described the basic logic of such free-for-all systems in the 1600s, and it hasn't changed much since.

The bottom line, in such a world, will be a ceaseless competition to build stronger forms of AI – for all the players will know that AI technologies hold the key to geopolitical power. Every national government and every major informatics corporation will have teams of researchers frantically seeking new breakthroughs. In such a condition, safety concerns inevitably tend to be relegated to a lower priority: what matters most is staying ahead of the other racers (or at least, not falling too far behind).

In this sense, AI is a quite different type of technology from nuclear weapons when it comes to its impact on geopolitics. Once you have a certain number of nukes, you get diminishing returns if you insist on building more and more: going from five thousand warheads to twenty thousand warheads does not buy you four times greater leverage in world affairs.[22] But AI doesn't work that way. With AI, the rising gradient of machine capabilities *does* translate to proportionate increases in geopolitical power: as your nation or corporation progresses from zero AI to narrow AI to AGI to ASI, your influence over world affairs will continue to go up dramatically with each advance. (To be sure, after the technological level reaches ASI, all bets are off, particularly if it comes about via a hard takeoff.)

This is a recipe, sooner or later, for bad outcomes. In an ongoing, no-holds-barred race for ever-stronger AI, with all the key players focused on being first past the post, the likelihood of a hard takeoff stands at its highest. Whether such an intelligence explosion were unleashed deliberately or accidentally, the risks are equally extreme. And even if an intelligence explosion is somehow avoided, year after year, sooner or later one key player will stand a good chance of gaining a decisive technological advantage – which could lead to the emergence of a global hegemon with all its inherently undemocratic and unpalatable features. There's no way to sugar-coat this: if the international self-help system continues to prevail unaltered through the coming century, we are probably in for big trouble.

The good news is that a quite different pathway also lies open to us. Humankind can gradually piece together a different kind of global system in which the Hobbesian logic of all against all no longer applies. This would still be a multipolar order with no

AGI monopoly, but in this scenario the world's peoples would incrementally ramp up cooperative governance to much higher levels.

Vignette 8: Cooperative coexistence with superintelligent machines (2119)

I was born in the wrong century. That's what everyone tells me – my wife and son included. I should have been a cowboy in the Wild West, they say. Or a gladiator in ancient Rome. Or a Spanish conquistador.

Then they laugh and walk away.

I feel like taking off my shoe and throwing it at them. But they have a point. I just don't fit in. Our household Synthia, Bernie, puts it more gently: "The problem seems to be, Bruce, that you simply embody the wrong virtues for this day and age."

I want to make my own damn coffee when I get up in the morning. I want to make my own freaking mistakes as I try to figure out the stock market. I want to learn things on my own, not have some bland-voiced robot explain everything condescendingly to me in two-syllable words. And don't try to tell me a machine can't be condescending. I know better.

When I was 26 I got so fed up with it all, I joined a neo-hippie commune in southern Oregon, near the Deschutes forest. No robots, no bioenhancements, no communication implants: just farming and chickens and goats and getting up in the morning with the rising sun. It was fantastic. For a while. Then I started getting bored with the same ole, same ole. I wanted new challenges.

That was when she came along. Jenny, my new challenge. She was working the white-water rafts on the Rogue river, taking a summer break from college before her senior year. I met her in the tavern in Agness, where the rafting people congregated in the evenings. The server bot seats me next to her at the bar and I order an old-fashioned IPA and she looks over at me and grins. "I just won a bet with myself," she says. I look over at her. She's got a really pretty smile. "Oh?" I say. She nods: "You didn't look like a synthetic brew kind of guy."

Two years later we were married and Jenny gave birth to our kid, Robert. Neither of us needed a job because the Basic Income was more than enough for our needs. But I strongly believed a man or woman had to do something useful for the world, not just take handouts from the goddamn machines. I still believe it today.

Jenny tells me I'm a grump. I'm ungrateful. I'm a pain in the butt. Worst part is, she says it lovingly, reaching over to pull a strand of hair back from my face, like these are my most delightful personal qualities.

I eventually got a job working for the United Nations. My thinking was: go directly into the belly of the beast. We've lived for two-year stints in Copenhagen, Lagos, Buenos Aires, St. Petersburg, Denver, and Shanghai – some of the core directorates where the U.N. does its daily business.

I like the travel, and so does Jenny. We meet all kinds of people. Some have become good friends.

My job is actually pretty cool. I'm an investigator for the U.N.'s AI Security Agency; our core mission is preventing anyone on the planet from building a superintelligent machine outside the strict guidelines set by the Security Council. Despite

the ferocious penalties, we still keep catching people, year after year. I guess it's human nature: some people just can't resist. They want to be the ones who say, "We did it first. We built the first exponentially self-modifying AI." They tend to think they're the one group of special geniuses who can pull it off safely and successfully. That's what they insist at their trials, one group after another, year by year. And off they go into the slammer, life without parole.

Some of those rogue teams have actually come closer to unleashing an intelligence explosion than the public realizes. And I have to admit: if it weren't for the Synthias, we'd never succeed at our mission. It's always the Synthias who first detect the telltale pattern of behaviors that tips us off: a particular succession of hardware purchases, a string of carefully-concealed financial transactions, an unexplained growth in bandwidth use on the Web; certain kinds of algorithms circulating on the dark Web; certain word patterns recurring in the encrypted communications we're constantly monitoring. Only the network of Synthias working together can keep up with such a complex monitoring task. I guess it takes an AI to stop an AI. One of the little ironies of our time.

The global network of Synthias is legally owned by all humankind, and governed under the collective trusteeship of the U.N. Just like the nukes. That's the only way to keep a single nation or corporation from cornering the market on AGI and dominating everyone else. The power over the machines – and the power the machines create – have to be shared by everybody. It's the only way to prevent a planetary dictatorship from forming.

Sometimes it still feels like a dictatorship anyway. We humans are nominally in charge – each of our national delegations rotating through the various tiers of the Security Council, sharing power. But who are we kidding? The machines are so much smarter than we are – they're the ones who come up with most of the ideas we end up adopting. Our laws, economic system, ecological practices, military decisions, industrial policies – it's all devised by the network of Synthias in the end. I've been present at some of the Security Council meetings where it happens: the humans listen to the machines, the humans deliberate, the humans debate, and at some point we all just look at each other and say, "It makes sense. We should do what they suggest."

I used to find this demoralizing. Especially when I was younger. But now I just accept it for what it is: we've created beings that are smarter than us, so we may as well benefit from their smartness.

Don't get me wrong, the world still has serious problems. But today they're no longer problems of survival, like they were fifty years back – nuclear war, pandemic, ecological crisis, stuff like that. We've figured out ways to work together in dealing with those.

Today our toughest problems are about identity. Who are we? Are we becoming like household pets of the machines, the way some religious activists claim? Is homo sapiens *going to fragment into a multitude of successor species, now that bioenhancement technologies allow us to alter our bodies and minds in any way we wish? Are we becoming increasingly like machines ourselves?*

Jenny and I argue about this sometimes. Especially when it comes to choosing bioenhancements for ourselves and our son Robert. OK, we can both agree about the genetic tweak for resistance to cancer. Same for the drugs we use to increase health

span for the three of us: I turned 58 last month, and I feel like a twenty-year-old. Jenny still looks as youthful as when I first met her.

But what about the brain implant for direct Web connection? The epigenetic tweaks that let you fine-tune your emotions moment by moment? The skull caps you put on to share intimate memories more directly with each other?

Last night at the dinner table I told her: this whole thing just doesn't feel right. It's like I'm becoming a kind of semi-designed product. I feel like I'm a marionette, and I'm the puppeteer at the same time.

She smiled and reached across to pour me more wine. "What do you want to do? Go back to the Deschutes forest?"

I scowled down at my plate.

"Look at the big picture, Bruce. The world's not as bleak as you're always making it out to be. These questions, these dilemmas – this is a good set of problems to have."

I looked across at her for a moment. Her smile, the depth in her eyes. I reached for my wine glass and took a slow sip.

And I thought to myself: I'm still getting up tomorrow morning and making my own coffee. Even though Bernie's tastes better.

[1] Carl Sagan, *Pale Blue Dot* (Ballantine, 1997), 305-6. I have altered the punctuation and line spacing of the quoted paragraph for greater effect.

Notes to Chapter One

[2] See the discussion in chapters 2 and 4.
[3] See the discussion in Chapter Four.
[4] For a systematic effort to envision such a world, populated by advanced robots built via the design pathway of whole-brain emulation, see Hanson, *The Age of Em.*
[5] The most comprehensive discussions of this scenario are Bostrom, *Superintelligence*, especially chs. 4, 11, and 14; and Tegmark, *Life 3.0*, chs. 3-5. See also Khatchadourian, "The Doomsday Invention: Will artificial intelligence bring us utopia or destruction?"
[6] Bostrom, *Superintelligence*, chs. 4, 11, 14; Tegmark, *Life 3.0*, chs. 3-5.
[7] Bostrom, *Superintelligence*, ch. 5.
[8] Cade Metz, "Inside OpenAI, Elon Musk's Wild Plan to Set Artificial Intelligence Free," *Wired* (April 27, 2016); Theo Priestley, "Does Elon Musk And OpenAI Want To Democratise Or Sanitise Artificial Intelligence?" *Forbes* (Dec. 13, 2015).
[9] The list of sponsors for OpenAI is here: https://openai.com/about/#sponsors
[10] See the rationale for creating OpenAI on the following page of the organization's website: https://blog.openai.com/introducing-openai/ The mission statement of OpenAI is at: https://openai.com/about/#mission
[11] This quotation is taken from the charter of OpenAI, released on April 9, 2018: https://blog.openai.com/openai-charter/
[12] The physicist Max Tegmark describes Musk's commitment to AI safety in Tegmark, *Life 3.0*, 321-27.
[13] Milton Mayer, *They Thought They Were Free: The Germans, 1933-45* (U. of Chicago Press, 2017); Konrad Jarausch, *Broken Lives: How Ordinary Germans Experienced the Twentieth Century* (Princeton U. Press, 2018); Hanna Arendt, *The Origins of Totalitarianism* (Harcourt, Brace, Jovanovich, 1973); Remnick, *Lenin's Tomb*; Moshe Lewin, *The Making of the Soviet System* (Pantheon, 1985); Jeremy Brown and Matthew Johnson, *Maoism at the Grassroots: Everyday Life in China's Era of High Socialism* (Harvard U. Press, 2015).
[14] Christopher Browning, *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland* (Harper, 2017); Norman Naimark, *Genocide: A World History* (Rwanda Stanley Milgram, *Obedience to Authority: An Experimental View* (Harper & Row, 1974). Victoria J. Barnett, *Bystanders: Conscience and Complicity During the Holocaust* (Praeger, 2000); Inga Clendinnen, *Reading the Holocaust* (Cambridge U. Press, 1999); Peter Haas, *Morality After Auschwitz: The Radical Challenge of the Nazi Ethic* (Fortress Press, 1988); Raul Hilberg, *Perpetrators, Victims, Bystanders: The Jewish Catastrophe, 1933-1945* (HarperCollins, 1992); Michael Marrus, *The Holocaust in History* (U. Press of New England, 1987); Ervin Staub, *The Roots of Evil: The Origins of Genocide and Other Group Violence* (Cambridge U. Press, 1989); Philip Gourevitch, *We Wish To Inform You That Tomorrow We Will Be Killed With Our Families: Stories From Rwanda* (Picador, 1999).
[15] George Orwell, *1984* (Signet, 1950); Aldous Huxley, *Brave New World* (Harper, 2006).
[16] Bess, *Our Grandchildren Redesigned*, chs. 10-12.
[17] *Star Trek: The Next Generation* (Paramount, 1987-94).
[18] *New York Times*, "Russell on Preventive War" (Dec. 3, 1961); Ronald Clark, *The Life of Bertrand Russell* (Bloomsbury, 2012), ch. 19.
[19] Combs, *The History of American Foreign Policy From 1895*; Kaufman, *A Concise History of U.S. Foreign Policy*; Peterson, *American Foreign Policy*.
[20] Brown, *Seven Years That Changed the World*; Horvath, *The Legacy of Soviet Dissent*; Service, *The End of the Cold War*; Sebestyen, *Revolution 1989*; Sarotte, *1989*; Grachev, *Gorbachev's Gamble*; Remnick, *Lenin's Tomb*.

[21] Lee, *AI Superpowers*; Brynjolfsson and McAfee, *The Second Machine Age*; Philippe Van Parijs and Yannick Vanderborght, *Basic Income: A Radical Proposal for a Free Society and a Sane Economy* (Harvard U. Press, 2017).
[22] Schelling, *Arms and Influence.*

Notes to Chapter Eighteen