Hera
THE BUDDHA

A WORK OF UTOPIAN FICTION BY

WILLIAM LEISS

BOOK THREE OF THE HERASAGA

The Herasaga

Figure 1 Yucca brevifolia in bloom, Joshua Tree National Park, California (Photo: W. Leiss)

HERA THE BUDDHA

A WORK OF UTOPIAN FICTION

WILLIAM LEISS

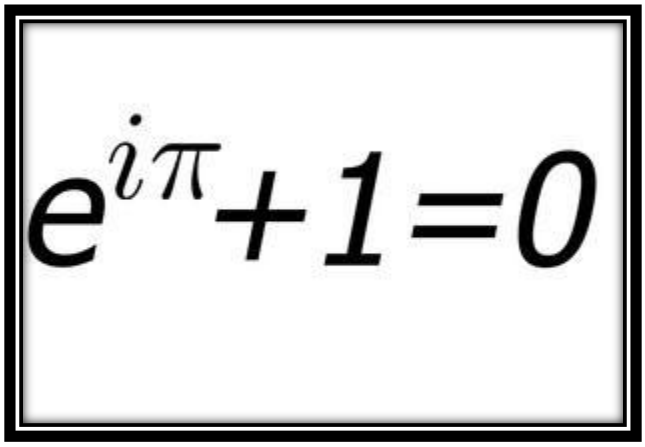

Figure 2 Euler's Identity

A CANGRANDE BOOK

A CANGRANDE BOOK
Ontario, Canada
william.leiss@gmail.com // www.leiss.ca

This is a work of fiction. Names, characters, places, and incidents are products of the author's imagination or are used fictitiously. Any resemblance to actual persons, living or dead, is entirely coincidental.

Library and Archives Canada Cataloguing in Publication
Leiss, William, 1939-
Hera the Buddha: A Work of Utopian Fiction/
William Leiss.
(The Herasaga ; bk. 3)
"A Cangrande Book".
ISBN 978-0-9738283-2-0

I. Title II. Series: Leiss, William, 1939- Herasaga ; bk. 3.

COVER ARTWORK: ALEX COLVILLE (CANADIAN 1920-2013),
MOON AND COW (1963), OIL AND SYNTHETIC RESIN ON HARDBOARD
PRIVATE COLLECTION, USA

COVER DESIGN BY HYDESMITH COMMUNICATIONS, WINNIPEG

for HEIDEMARIE and THE DAUGHTER

EPIGRAPHS

What happens when machines become more intelligent than humans? One view is that this event will be followed by an explosion to ever-greater levels of intelligence, as each generation of machines creates more intelligent machines in turn. This intelligence explosion is now often known as the "singularity." …. If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet.

David Chalmers (2010)

As if somehow intelligence was the thing that mattered and not the quality of human experience. I think if we replaced ourselves with machines that as far as we know would have no conscious existence, no matter how many amazing things they invented, I think that would be the biggest possible tragedy. There are people who believe that if the machines are more intelligent than we are, then they should just have the planet and we should go away. Then there are people who say, 'Well, we'll upload ourselves into the machines, so we'll still have consciousness but we'll be machines.' Which I would find, well, completely implausible.

Stuart Russell (2017)

We are the first species capable of self-annihilation. Elon Musk (2017)

If you want a picture of A.I. gone wrong, don't imagine marching humanoid robots with glowing red eyes. Imagine tiny invisible synthetic bacteria made of diamond, with tiny onboard computers, hiding inside your bloodstream and everyone else's. And then, simultaneously, they release one microgram of botulinum toxin. Everyone just falls over dead. Only it won't actually happen like that. It's impossible for me to predict exactly how we'd lose, because the A.I. will be smarter than I am. When you're building something smarter than you, you have to get it right on the first try.

Eliezer Yudkowsky (2017)

[W]e need not worry about the forecast that, in the near future, a "really smart" digital computer/machine will supplant human nature or intelligence. In all likelihood, this day will never come because, in a more-than-convenient arrangement, our most intimate neural riddles seem to have been properly copyright-protected by the very evolutionary history that generated our brains, as well as the very complex emergent properties that make it tick. As such, neither evolution nor neurobiological complexity can be effectively simulated by digital computers and their limited logic.

Miguel Nicolelis (2014)

List of Figures

Table of Contents

Prologue and Retrospective

THIS IS BOOK THREE in the trilogy I have dubbed *The Herasaga*, being the sequel to the earlier volumes in this series – the first, entitled *Hera, or Empathy,* and the second, *The Priesthood of Science*. I must confess at once that it is I alone who am responsible for associating my aunt Hera with the revered figurehead of the notable Eastern religion which took his name. For the eponymous figure in question undoubtedly would regard entitling a book *Hera the Buddha* as at best a presumptuous and at worst a disrespectful act. Gautama Buddha is revered as the one who experienced a sudden awakening, representing an escape from the endless cycle of ordinary existence and an insight into the hidden truth of being, and his followers have regarded his life and teaching as representing a radical break in human existence. I confess that the analogy I have constructed here is based on a simple, even simpleminded idea, namely, that of a human figure who sees his or her life as being positioned in a transitional state between one historical era and its radically-different successor.

Hera understands her own time as an essential part of a long period which witnessed the passage from a collective mind dominated by religion to one where modern science came to rule the interpretation of both nature and human nature. This process she saw as a belated and still-fragile fulfillment of the promises made in the 18th-century French Enlightenment. For beginning in the seventeenth and eighteenth centuries in Europe, the great founders of the new science of nature saw science's role in terms of a broader mission, whereby an approach using the experimental method and evidence-based reasoning would replace "superstition" as the basis of social organization.

The champions of the new science and Enlightenment were well aware that they faced a determined and resourceful foe in the organized religions. But by the end of the twentieth century they had won the main battle hands-down. In most advanced

industrial economies by that time (the United States being the notable exception) religion had been relegated to the sphere of private life, and both the social order and its legal structure were thoroughly secularized. From that time onwards there was simply no credible challenge that theologians could mount against the conceptions underlying the sciences of nature in physics, cosmology, chemistry, and biology and in the applied fields such as medicine. Where the science-based explanations on offer were still incomplete, no one seriously believed that anything other than additional applications of the same scientific approach would remedy the deficiencies. And although beyond a purely superficial level most citizens would have been unable to articulate any of the basic concepts of these sciences, they all had complete faith in the mission of science itself, believing that only by this means would the newest answers to their needs, whether for faster electronic gadgets or improved cancer therapies or a hundred other advantages of this type, be delivered to them.

And then in the collective unconscious of the scientific community the bold vision sketched early in the seventeenth century by one of the founding fathers of the new sciences of nature, Francis Bacon, was revived. The new method, Bacon maintained, would not only tell the truth about nature's workings, for the first time, but would also give rise to an unending stream of new practical applications to improve the conditions of human life. It would eventually, he predicted, lead to "the effecting of all things possible" by reversing the terms of the relation between nature and humanity. Whereas earlier humans had experienced passively the power of natural forces, suffering great harms in the process, the new sciences gradually would allow humans to exercise more and more complete power over nature – in the sense that the understanding of how natural processes work would, by allowing humans to follow the previously hidden cause-and-effect chains, permit humans to "lead" nature to their own chosen ends for the betterment of life. For example, understanding the mechanisms of gene expression allowed us to change the genomes of plants and animals, including ourselves, of course. Bacon set out this vision in 1625, and subsequent events proved him right. He anticipated the results, and would not have been surprised by them.

Buried in his dream was the idea that humans would stand in their Creator's place as absolute masters over nature, able to turn to their own advantage every capacity for action that could be discovered in nature's powers.

The idea of absolute power – in the sense of complete power over nature generally and over the fate of humanity in particular – is originally a product of Western monotheism, as developed first by Judaism and modified later by Christianity and Islam. This power is exemplified in the act of creating nature *ex nihilo*, "out of nothing," and of course what is created by God's hand can just as easily be obliterated again in its entirety. And yet in these three monotheisms the vast diorama representing the starry heavens and the life-sustaining earth is merely a stage-prop for the real and exclusive business of creation, namely, the human drama: What happens to the rest of nature after its coming-into-being is quite irrelevant, meriting no further comment in the sacred texts. On the other hand, in those monotheistic texts absolute power *over the fate of humans* is vividly displayed.

Hera thinks that what transpired under the monotheistic representation of the exercise of absolute power represents a cautionary tale for the rest of us. In short, she opined, *the tradition of Abrahamic monotheisms shows us that there is something about the possession of absolute power – meaning a power so complete that any action which can be willed at any moment is available for deployment – that unhinges the mind*. On the surface it appears to us as a source of benevolent and intriguing capacities, conferring rewarding experiences such as eternal life (in the religious tradition) and time-travel or genetically-enhanced brain functions (in scientific fantasy) on its lucky beneficiaries.

But what strikes us so strongly in the religious version is the sheer intensity of the destructive and limitless *rage* that accompanies the wielding of irresistible power. In many of the holy books of the three Abrahamic religions, and in the countless commentaries on them by later authorities, this rage, directed at both believers and unbelievers alike, is ascribed to the Supreme Deity. It seemed to Hera that science's champions needed to reflect deeply on the ultimate scope and potential uses of the

powers that they, now standing in the Creator's stead, were prepared to bestow upon those self-same intemperate and weak-willed creatures for their enjoyment. In her mind, there is no possibility that the scientific community as a whole, as the creators of these powers, can evade their ultimate responsibility for the uses to which they are put. The set of reflections she prepared on this subject, and which I have compiled in the following pages, provide the unifying thread of the present volume.

Here I offer a brief *précis* of the two earlier volumes in this series, for those who have not yet perused them, beginning with the first, *Hera, or Empathy*. The story line features a band of sisters, twelve in all, who are the natural offspring of two parents, Ina Sujana and Franklin Stone, conceived by means of *in vitro* fertilization using previously-frozen eggs and sperm and subsequently brought to term in the wombs of surrogate Indonesian mothers. Their parents, both famous neuroscientists, had planned in advance to carry out germline genetic modifications on the embryos—alterations in certain genes that would be inherited by any future offspring of their daughters. The modifications targeted genes responsible for the development of specific regions in the brain's prefrontal cortex, in particular, the mysterious complex of neurological functions that is referred to as our sense of empathy. What they were hoping to achieve by "upregulating" these genes was to create a type of human being who would be strongly motivated to promote the common welfare of humankind.

All the sisters were born within about ten days of each other in the month of September 2014; Hera, the first-born of the lot, assumed a leadership role in the band from an early age. She was fifteen when she first learned what her parents had done to them, and what her father's larger plans were (their mother had died some years before they were born and at the time they were separated from their father). After the sisters were reunited with their father she began questioning him about his motivation in engineering his daughters, a topic charged with much emotion in itself, but made more so by her knowledge that he was determined to repeat the procedure on a far larger collection of embryos.

Her increasingly fractious colloquies with her father always left her unsatisfied. She simply couldn't dent the armor of his self-confidence in the triumph of modern science: We know now how nature works, we are the new masters of creation, we have a duty to advance our understanding continuously and also to turn our knowledge to the betterment of humankind, a task that includes the re-engineering of genomes, human, plant, and other animal, to relieve suffering and enhance our capacities. Yes, inevitably we will make mistakes, but we will learn from them as well, and we will utilize the powers bestowed on us by science to mitigate and remedy those mistakes.

Then one day there came from the biologists among the sisters the revelation that they might be hybrids—a new variant of the species *Homo* that might not be able to breed successfully with other humans. Eureka! A new species – or subspecies at least? The full implications of this awareness hit her like a thunderbolt. Somehow, her father had done this to them, unbeknownst to him, and unintentionally, of course. He was preparing to do it again, on a thousand embryos created from eggs and sperm donated by talented couples representing diverse genetic subtypes from every corner of the globe. Well then, why not? Let's make both sexes this time, on the assumption that they'll be fertile *inter se* (among themselves), and then we'll branch off from other humans and travel down our own path of evolutionary development. And so she rolled the dice and the great gamble commenced.

But how to divine what is the right path? After all, evolution – at least until now – is what happens by chance to any one species, willy-nilly, in the encounter with its competitors, not a set of conscious choices. There are no precedents in the written record – save perhaps the one found in Mary Shelley's *Frankenstein.* The nameless and tragic being portrayed in that famous story, whom his creator Frankenstein had been wont to call "the demon," was in fact a new species, the first of its kind. In the creature's demand that the scientist should acknowledge the duty to extend his own mission to its logical end, by fashioning for him a female of his own kind, Frankenstein realized to his horror that the venture on which he had embarked now threatened to become another Adam-and-Eve story. True, this had not been his original intention: Do you see how the

situation of Hera's father mimics Dr. Frankenstein's? What the creature's demand had done was to expose the unpleasant truth that this bold scientist had simply failed to think through the full implications of his experiment with life.

Hera's reflections on the comparison between the twelve sisters and Frankenstein's solitary creation ended thus: Far more than he was, we are products of modern science, or more fully, products of science's penetration of the processes of biological life and evolution and its subsequent mastery over those processes. When she got there she said to herself, in effect: This is how and why I was created, namely, as an integral moment in science's will to grasp the secrets of organic life and to use that knowledge in new technologies. This act of human reason and will defines the uniqueness of my own being as a conscious entity on earth. I exist as such: The deed has been done, I am aware of how and why it was done, and I must accept my situation. And I must also therefore "appropriate," or fully absorb and internalize, *my* awareness of this act, considered as an intentional exercise of uniquely human powers, *as well as the feelings which this process of absorption induce within me*.

Having placed her coming-into-being in the context that made the most sense to her, Hera concluded: "I am terribly troubled by the process of my creation. What troubles me above all is the disjunction I see between the careful rationality of the scientific mode of reasoning, on the one hand, and the arbitrariness of the will that chose to carry out this specific act of genesis, on the other. Under another type of protocol made feasible by science's mastery over biological processes, I might have been brought to life as a malevolent entity, designed to be an instrument in some despicable plan to sow horror and misery among humans or to subject them, on my creator's behalf, to bitter and eternal oppression. Using the words in which Frankenstein expresses his fears about what the demon and his mate might represent, once their kind had begun to reproduce, I myself could have been fashioned into a being 'who might make the very existence of the species of man a condition precarious and full of terror.' This thought is intolerable to me."

Why intolerable? Because, so far as her new species was concerned, science was not just the progenitor of desirable technologies. To them it would always and forever also have, quite literally, an *existential* meaning, since it had shaped who they were meant to be as well as how they had come to be. As creatures designed by the scientific enterprise itself, they stood on its shoulders and saw the surrounding world of nature through its eyes and with its mental constructs. She even hoped they would earn the right to brand themselves with a unique scientific name: *Homo sapiens scientificus.* But the collapse of industrial societies around the world, and the ensuring social chaos and unrelenting bloodshed as billions of impoverished people fought over a dwindling pile of resources, meant that they might never get the chance to show what they were capable of unless they could wall themselves off from the general mayhem.

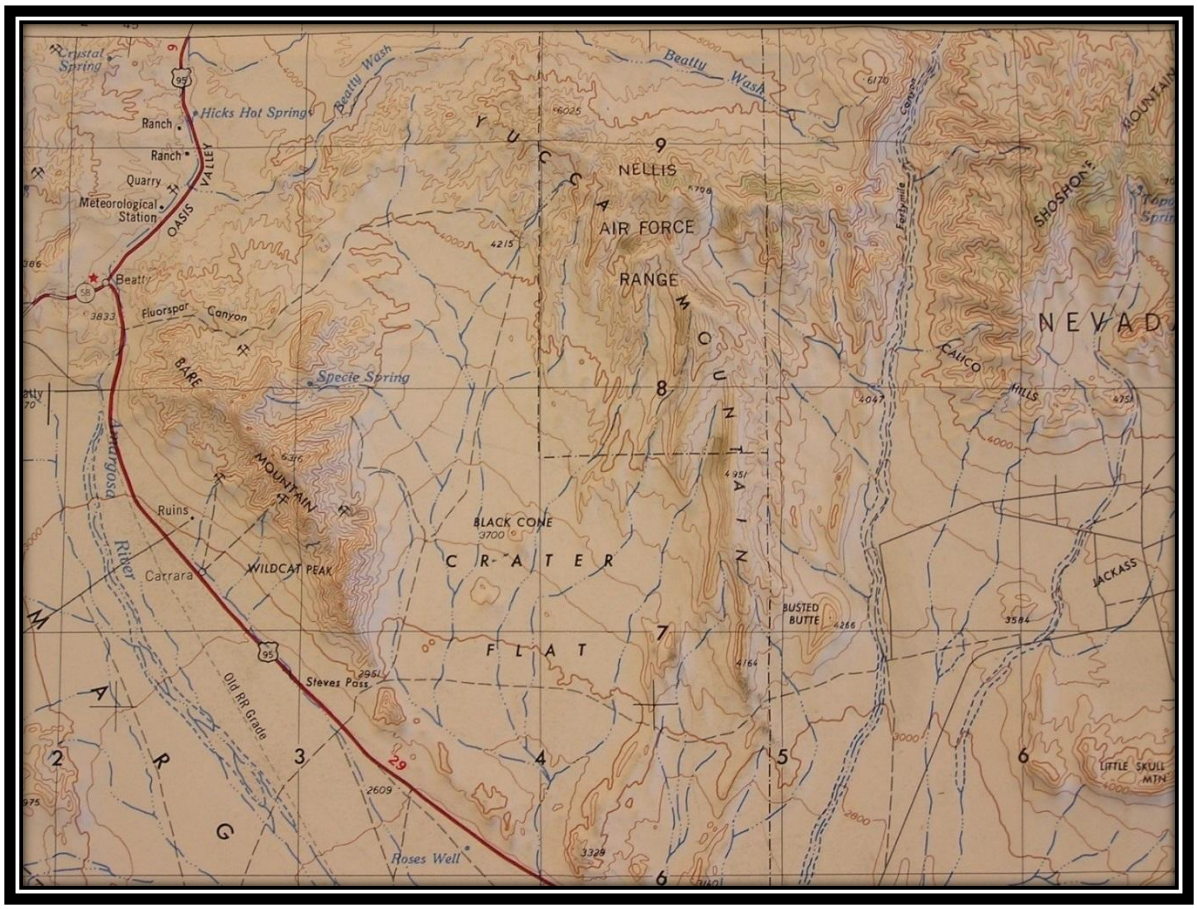

Figure 3 Crater Flat, in the Mojave Desert, Southwestern Nevada (Map Photo: W. Leiss)

So, she went looking for an isolated and easily-defendable permanent site for her tribe, and found it in the middle of the Mojave Desert, in a place called Crater Flat, nestled between Bare Mountain and Yucca Mountain (the site of a nuclear waste repository) in southwestern Nevada, roughly about 80 miles northwest of Las Vegas Valley. She obtained a complete library of world science on videodisc and salted it away, along with as much up-to-date scientific instrumentation as she could scavenge, inside secure tunnels next to the mountain's cache of nuclear waste. And, some years before the date when this present volume commences, fearing that the accelerating social chaos around the globe might among other things bring down in ruins, one by one, the world's premier scientific research institutes, she had asked her biologist sister Gaia to begin organizing a vast rescue mission in order to gather up leading scientists and their families, wherever they could be found, and install them in safe quarters at the private university they had founded in the then-vacant city of Las Vegas.

The rest of the story is told in *The Priesthood of Science.* As the sequel opens, in March of 2064, the finishing touches are being put on the core facilities at Yucca Settlement, all financed out of the enormous resources accumulated in the sisters' private charitable foundation. It is self-sufficient in food, water and necessities; increasingly isolated as towns, cities, and ranches within a 100-mile radius become deserted due to persistent and severe drought; well-defended by its own formidable weapons emplacements; and further protected by the fighter-plane squadrons located nearby at Nellis Air Force Base, our contractual partner in the lease arrangement for the land we occupy. And, most important of all, a little more than a thousand talented children live in the Settlement, the second generation of Hera's kind, who are about to turn eighteen.

In their desert hideaway, her people are well-concealed and well-provisioned. Moreover, Hera and her sisters have succeeded in rescuing thousands of leading scientists from around the world, who had witnessed their research institutes being destroyed by marauding bands, installing them and their families in secure new facilities

equipped with the latest scientific gear, emplaced in a network of abandoned buildings in Las Vegas. They are also, through communications networks linked by satellites in earth orbit, in close touch with a small group of like-minded entrepreneurs elsewhere in the world, who are also organizing refuges where scientists and their families can carry out their activities in safety. They shared with the sisters who run the Yucca Settlement a simple belief that the coming of modern science, grounded in a rigorous method of evidence-based reasoning, marked a watershed in human development that must be preserved amidst the global social disintegration. If lost, it might possibly never be found again, for it arose by chance; and in the opinion of its protectors that would be the worst tragedy of all.

Like the rest of these entrepreneurs, Hera and her sisters were well aware that many of the inventions based on the discoveries of the modern sciences had contributed to the social collapse occurring around the world. It was during the formative period of the new science of atomic physics, in the second and third decades of the twentieth century, that the ultimate price humanity might have to pay for science's astonishing advances became all too clear. For that new science had achieved its early successes during the rise of fascist barbarism, with its dreams of an entire world subjugated to a permanent regime of terror established by a self-proclaimed master race, an intrinsic absurdity.

Over a period of a little more than a decade the new physics evolved from the first full description of the internal subatomic structure of matter to the idea of a "wonder-weapon" of previously unimaginable destructiveness. Only some good luck ensured that this weapon was not available for Hitler's arsenal. But only two decades after it was used to end World War II, at the height of the Cold War, there was enough destructive power in the hands of opposing nations armed with nuclear bombs to reduce all the glories of human civilization to a radioactive ash-heap.

Since that time other similar threats had arisen, especially those based on the idea that one or more naturally-occurring pathogens, such as a virus or bacterium, could be deliberately engineered with the help of molecular biology so as to avoid all defenses

against it and produce a pandemic whose human victims might total hundreds of millions or even billions. Another threat was even more sinister, in a way, because it was presented as something so benign and indeed so universally advantageous as to have no downside at all: The engineering of new traits into the human genome, in the germline, so that the modified genes could be passed down to future generations. Who could object to the permanent elimination of unwanted, debilitating conditions such as inherited diseases? Who would oppose the introduction of enhancements across the whole spectrum of human physical and mental capabilities? What person was there who could find no use for greater endurance, more capacious memory, and better performance in what is known as our brain's "executive" functions? What if science told us that we could even integrate the astonishing information-processing powers of silicon-based computers seamlessly with the human brain's neural tissue?

In all these realms and others scientific genius was applied, the obstacles fell, and the products were delivered to a grateful populace. So far, so good. But not everyone outside the scientific community had the same end-uses in mind for these remarkable innovations, nor the same level of access to them. In a world made up of fractious, belligerent and highly unequal nations and social classes, all such innovations confer competitive advantages to some persons over others. Those who are in a position to appropriate them first have very little interest in seeing them universally shared. To be sure, this had always been the case.

But the paradox of affluence accentuated, rather than diminished, the maldistribution of new resources and new possibilities for human development. This paradox dictates that the increase of wealth gives rise, in most cases, to a powerful urge to appropriate still more, without limit, while also ensuring that the underprivileged masses have no means of upsetting the applecart. How much more advantageous would it be if such disparities could be cemented forever into inherited genetic endowments? The answer is obvious.

Over the course of four centuries the enterprise of modern science had freed itself from all external constraints, represented by religion and the state, so far as its

distinctive methods were concerned. Increasingly the practice of science was autonomous, a self-governing system, and it flourished in this glorious freedom. The pace of discovery accelerated under the conditions of this ever-enlarging autonomy. Already by the nineteenth century organized religion had been sidelined as a potentially interfering and controlling overseer. Over the course of the twentieth century the nation-state was simply bought off: Scientists dumped the practical results of their discoveries – radically new materials, processes, devices, and entire realms of reality (e.g., the nanosphere) – into the public domain, and said, in effect: "Help yourselves. Enjoy. But don't interfere, just come up with the research budgets but otherwise leave us alone, or else the spigot will be shut off."

Governments happily appropriated some of the innovations for fancier weaponry and means of more efficient population control, such as surveillance. The rest could be allowed to find its way into the economy, to make the wealthy wildly richer and the rest stuffed with cheap protein and distracted by their electronic gadgets, the tried-and-true formula of bread and circuses, now technologically-advanced. Everybody was happy, except the desperately poor, who had no power to alter the arrangement.

With every new phase of scientific discovery, the operational powers of human agents – their capacity to intervene in natural systems and with new technologies to alter the conditions of human experience – grew qualitatively and quantitatively. And yet other, quite decisive, aspects of human history remained curiously unaffected by what modern science was doing. Militant organized monotheisms continued to flourish, and religious fanatics of one faith or another carried on slaughtering those who worshipped other gods or even other versions of the same deity, motivated, in the case of Islam, by a schism that had developed in the eighth century.

Old ethnic divisions continued to supply other justifications for regular bloodbaths: At the onset of the twenty-first century the Serbs in Kosovo were said to be taking revenge on their neighbors for something obscure that happened in the fourteenth. Old national entities, such as Japan and China, kept ancient animosities alive and well; newer ones such as India and Pakistan kept nuclear-tipped missiles aimed at each other

at the ready at all times. In the case of the former Soviet Union the world watched with interest as the state with the second-largest nuclear arsenal in the world, as well as a massive and highly-sophisticated bioweapons store, together sufficient to exterminate most of the human race several times over, slowly disintegrated into a novel political entity run by criminal gangs and nineteenth-century-era robber barons.

Hera and her comrades sought to understand these curious developments whereby scientists kept throwing the means for developing clever new devices with frightening destructive powers into the hands of madmen and fanatics whose motivations were grounded in long-ago times. And all the while, with lawyerly acumen, these scientists would insert the following implicit proviso into all of their contracts with society and the state: "The recipients of these new devices hereby absolve the inventors thereof of all responsibility for whatever uses the latter find for them, and for all the consequences of those uses." How very convenient, and how very absurd. However absurd, it was an unquestioned article of faith among academic scientists, whose common practice was to ignore the issue completely.

The Priesthood of Science includes three long dialogues between a few of the sisters, Hera, Gaia, and Athena, on the one side, and one of their distinguished recruits, Professor Abdullah al-Dini, on the other. The sisters made a sharp distinction between science's underlying mission of evidence-based reasoning and its practical applications, affirming that it was the former in which the true value of modern science lay. They introduced him to the proposition that the advanced tools furnished by science to society were simply too dangerous to be deposited so casually into the public domain, available there for anyone so disposed to pick them up and use them for nefarious purposes. They surprised him with the conclusion they drew from their diagnosis, to wit, that the conduct of science must be hidden away and concealed from the public for the time being, until times changed and the likelihood of misuse had been substantially reduced. They said that this was a logical conclusion derived from science's hard-won autonomy from societal oversight.

They appealed to Goethe's famous story, "The Sorcerer's Apprentice," and to the example of the illustrious Francis Bacon, the seventeenth-century champion of scientific method, who in his utopian fantasy, *New Atlantis* (1627), had imagined a future in which the scientific community exercised its autonomy by deciding which of the practical applications resulting from scientific discoveries should be revealed to society, for the benefits they would bestow, and which would remain hidden, known only to their fraternity. And most fervently of all, they insisted that scientists could not absolve themselves of all responsibility for the uses to which their discoveries might be put. The story I told there recounted how Dr. al-Dini responded to these propositions and the reasons he advanced for not accepting them.

At the core of those dialogues was the notion that the sisters had adopted from one of the great leaders of the eighteenth-century French Enlightenment, the Marquis de Condorcet. He had seen that there was another, vastly more important aspect of modern science's triumph, beyond the marvelous new devices and powers it would invent. That other contribution to the welfare of humankind was the promise that science's core method of evidence-based reasoning would diffuse through all the institutions and practices of civilization, driving out the superstition and irrational fear that until then was the main foundation of human behaviors. He referred to the period predating the emergence of modern science as the infancy of humankind, to be succeeded by a qualitatively new epoch. He had, in effect, posed this question for those who would come after him: What does it mean to actually live out our lives "in accordance with" the understanding of nature that modern science bequeaths to us?

Unfortunately, Condorcet had underestimated the staying power of the previous order as well as the length and difficulty of the path leading toward a radically new type of society. He had not imagined that science would continue to deliver its new powers to a human social order that remained locked into the old pattern of internecine strife, irrational ideologies, and increasing violence. He could never have imagined that, one day, a science-loving humanity would have constructed a theater in which thousands of nuclear-tipped ballistic missiles, carrying a total destructive power capable of wiping out

everything of value accumulated by all prior civilizations many times over, would exist for decades in a state of permanent readiness for use. It is unlikely that he could have imagined that the scientific community itself, having in full view before it the eagerness of humans to use every new power invented by science in the service of ancient hatreds and superstitions, would just carry on with the established program.

Nothing seemed able to persuade most members of the scientific community even to take seriously the risk of catastrophe to which they were contributing. They drew no useful lesson from the example of Germany in the 1920s, a nation that had quickly assumed leadership in scientific discovery and industrial application – and then, unforeseen by anyone, when that same nation became the scene of a new round of collective human madness. None of them reflected on the possibility that the ultimate outcome of that scenario could easily have been far worse than it was, had just a few of the key variables (such as the contest between Nazi Germany and the Soviet Union, or the construction of an atom bomb) turned out differently.

None of them expressed concerns as, a century later, the then leading scientific nation, the United States, approached the threshold of a similar type of political irrationality. Then again, why should they? From its humble origins in the eighteenth century, the enterprise of modern science had by the beginning of the twenty-first triumphed beyond all earlier expectations: never had it had such an immense funding base, such a huge pool of brilliant and dedicated researchers, or (with a few exceptions, such as climate science and evolution) such abject deference from a public utterly enamored of a form of human activity whose intellectual basis they could not hope to understand. And so, they began to approach the end-game, where qualitatively new powers were under development, such as artificially-constructed conscious minds, robots with emotive capacities, nanomachines, the genetic enhancement of brain functions, and many others. As the possibilities stretched into areas never before contemplated, excitement grew. For a long time, it seemed that the good times in the laboratories would never end.

But end they did, because the structure's foundations were rotten through and through. The vast edifice of a hyper-rational science teetered on the shoulders of an incorrigibly irrational social order prone to paroxysms of mindless violence and terrifying brutality. Two starkly different choices presented themselves. The first was to let science fall into the maelstrom it had unwittingly helped to create and watch it be consumed in the fires of hatred. The second was to snatch it away and keep its spirit alive while preventing it from contributing anything more to humanity's arsenal, until the day came when a different ethos, firmly committed to the scientific spirit, prevailed in social relations. Hera and her sisters, and their like-minded colleagues around the world, had monitored these developments over decades. When they were in a position to intervene – in favor of the second alternative – they did so.

And while it was hidden away, the scientists who continued to serve their discipline would, Hera guessed, overcome their fascination with operational powers and seek only to deepen their understanding of nature itself and of ourselves as natural beings. She penned her thoughts on the pursuit of absolute power in the reflections collected in this volume in the hope that they would be helpful in this regard.

Marco Sujana
Director, Facilities & Operations
Yucca Settlement
Yucca Mountain, Nevada
CE 2070

Section One. The Mind Unhinged: Modernity and its Discontents

Chapter 1: The Rupture in Historical Time in the Modern West

AFTER THE COLLAPSE OF THE ROMAN EMPIRE, the three Abrahamic religions – Judaism, Christianity, and Islam – dominated Western civilization from the fifth to the eighteenth centuries CE. One of their salient features is a specific and deliberately backward-looking perspective. For the Pentateuch (the Five Books of Moses or the Torah), the gaze is on the seven days of creation story, which marks the beginning of time for humankind; for Christianity's New Testament, it is the birth of Christ in 33 BCE; for the Qur'an, it is the dictation of Islam's holy book to the Prophet Mohammed during the seventh century CE. Even though for the latter two, Christianity and Islam, there is an ominous foretelling of a future catastrophe known as the End Times or the Day of Judgement, the lineaments of that world-destroying episode were all unalterably set in motion by events in the far-distant past.

Despite a host of important technological innovations, such as in sea navigation, warfare, and agriculture, daily life for the common people had not changed or improved appreciably during that thirteen-hundred-year period. The Great Lisbon Earthquake of 1755, which also featured huge fires and a tsunami, killing tens of thousands, and which occurred on an important religious holiday, All Saints' Day (November 1), served as a reminder of humanity's helplessness before the traditional forces of untamed nature. Much more so, however, major catastrophes such as this one had always been interpreted in religious circles as signs of God's anger with persistent human transgressions against His commandments, again pulling attention back to an unchanging moral law and behavioral code set in stone so long ago.

But the Great Lisbon Earthquake also inspired one of that century's most famous tomes, Voltaire's *Candide* (1759), which mercilessly mocked this tradition. It was a sign, one of many, that during the eighteenth century a radically-different, comprehensive

challenge to established ways of thinking and acting was being mounted. We came to know this new intellectual force as the French Enlightenment. Although named for one nation, this revolutionary international movement also embraced Scots such as Adam Smith and David Hume, Englishmen such as Joseph Priestley, Germans such as Immanuel Kant and Johann Fichte, and French thinkers such as d'Alembert, Condorcet, Diderot, Voltaire, Rousseau, and Montesquieu; their signature collective work was the *Encyclopédie,* edited by Diderot and d'Alembert. All of them also looked back for inspiration to a set of somewhat earlier and equally influential English thinkers – Francis Bacon, Thomas Hobbes, John Locke, and Isaac Newton. And to the heroic Italian Galileo Galilei, of course, who as an old man had paid for his defense of his "new science" with house-arrest and the threat of torture by the Papacy's Tribunal of the Holy Office of the Inquisition.

In constructing their new world-view, Enlightenment thinkers had one powerful adversary in mind: Christianity, and more particularly, the Catholic Church. Their named their enemy "superstition," which they identified with organized religion, and to replace it they championed what we would today call evidence-based reasoning – or, more simply, modern science. In the eyes of many of them, most especially people like Condorcet, what was needed was for the type of reasoning implicit in the new natural sciences to gradually diffuse through modes of reasoning and behavior in the larger society. And in spawning both a continuing series of new technologies as well as industrialization, modern science eventually revolutionized not only thinking but everyday life to such a degree that it is hard to imagine anymore what life-conditions in the past actually represented for most people.

Gradually, however, science and its supporting mathematics became vastly more complex in purely intellectual terms. In general, the revolutionary insights achieved by the modern natural sciences unfolded in three historical stages – first chemistry, then physics, and finally biology. The eighteenth-century chemistry experiments by Antoine Lavoisier, Joseph Priestley and Henry Cavendish, as well as the earlier work of Robert Boyle, had already begun to have an impact on industrial production by the middle of

the nineteenth century. Physics was next, its development accelerating in the late nineteen-hundreds and then exploding in the first quarter of the twentieth century with atomic and subatomic theory, radioactivity, relativity, cosmology, and quantum mechanics. Biology followed, with great breakthroughs in evolution, molecular biology, and genetics, and at this point the deeper connections between these foundational natural-science disciplines also began to appear.

But as these sciences mapped out vast, previously-unknown dimensions of the natural world, their growing intellectual complexity gradually pushed all of them well beyond the capacities of ordinary human intelligence. In other words, the nature of the evidence they rely on to support competing theories and to verify experimental results defy easy explanation in terms of common-sense understanding. Beginning in the early 19th century there were frequent public lectures by scientists, but the most popular venues were such things as the theatrical demonstrations using electricity and magnetism, which fascinated audiences. Inevitably the natural sciences became over time less significant as a potential influence on popular thinking. Instead, what dazzled the popular imagination was the continuing outpouring of new technologies for the home, the workshop, and the factory. The machine age was upon us, first in mechanical form and later in electric and electronic.

The advantages of the new technologies of the machine age were so obvious, in terms of improving the general standard of living and reducing backbreaking labor, that the early forms of resistance to them – such as those of the Luddites – were easily overcome. When the consumer culture began to arrive, towards the end of the nineteenth century, bringing access to countless, helpful household devices for the majority of the population, continuous technological innovation became a key feature of everyday life. (This was the same period when advances in medicine meant that the treatments of physicians and hospitals started to become a net benefit rather than an incremental harm.) Once portable electronic devices arrived on the scene, daily life would never be the same again.

There is a strong interdependence between advances in modern science and in modern technology. This has been true since at least the time in the early seventeenth century when the new lenses developed by Dutch technicians were used in telescopes by Galileo and others to make startling discoveries about our solar system. This immensely productive interplay, where scientific discovery goes hand-in-hand with advances in measurement and detection instrumentation, continues to the present day: The extraordinarily complex technologies used in the Large Hadron Collider make possible the ongoing scientific discoveries in subatomic physics, most famously the experimental proof of the existence of the Higgs boson, which had been predicted by scientific theory. In a sense, this very fruitful interplay between science and technology goes as far back as their beginnings in ancient Greece. The best testimony to this is the remarkable "Antikythera" mechanism – called by some experts an analogue computer – dating to the period 150-100 BCE – an instrument with 30 meshing bronze gears, and designed for astronomical calculations; nothing rivalling its complexity would be constructed again until the 14th century in Europe. And yet the fateful nature of this interplay has been too little understood for far too long.

The thesis I wish to advance here is a simple one. It is this: The modern natural sciences, and more specifically their dependence on evidence-based reasoning, have been an unambiguous good for humanity and will remain so. But their closely-related phenomena, modern technologies and industrial society, are not. The future welfare of humanity depends on understanding this basic proposition as well as its practical consequences.

This thesis can be easily misunderstood, in part because the interdependence between science and technology is so obvious and significant. But that all or most technologies have a double aspect should be equally obvious. In their link to the sciences, technologies of observation and measurement are an indispensable part of humanity's insatiable desire *to know*. Their second aspect is the equally powerful drive *to act in the world*. The same or similar technologies are of course often found in both: For example, the new lenses crafted by Dutchmen in the 17th century, and used by

Galileo to make his revolutionary observations of the solar system, were also a huge aid to sea navigation. But the two uses are not logically linked in any respect. Quite apart from their employment in scientific knowing, technologies of action enter the social world – the "life-world" in Edmund Husserl's terminology – and are subordinated to the eternal quest for material prosperity, political power, and domination.

The first proposition – that the emergence of the modern natural sciences represents an ongoing, unambiguous good for humanity as a whole – was, of course, rejected by the Catholic Church in the seventeenth century. This is why Galileo was threatened with torture by the minions of the Inquisition. Scientific evidence remains unconvincing to many of those imbued with religious faith down to the present day. But this is largely a function of the needs of organized religions – fearing a reduction in their cash-flow prospects – to protect their own crass worldly interests, rather than the existential situation of individual believers, who are quite free to retain their beliefs no matter what the sciences say. And many billions of people apparently do just this; furthermore, they are likely to continue to do so, no matter what new scientific discoveries emerge in the future.

Modern science has never challenged religious faith directly (many competent scientists have been and are still religious believers of one sort or another). What it did challenge were various intellectual platforms – such as the opposition to heliocentric models – which various religious institutions had determined to be essential to the maintenance of their *secular* power. Gradually, many (but not all) of the church potentates realized that they did not have to fight this battle in order to ensure a flow of new believers across generations. Despite the emergence and ongoing successes of the modern natural sciences, various faiths in both East and West, including late novelties such as Scientology, have proved themselves to be the longest-running successful business proposition in world civilization.

Nevertheless, I hold onto the proposition that modern science is and will remain an unambiguous good for humanity. For it alone has – finally, after an immense collective effort and not a little courage – bestowed enlightenment on the human mind

with respect to our understanding of the *purely natural* processes that brought us humans into existence on our lovely planet. As the great scientist Laplace once said to the Emperor Napoleon, "I have no need of the hypothesis" which posits a creator-god. Natural chemical, physical, and biological processes, working together for Mother Earth, created first the eukaryotic cell and then all later life-forms, every one of which, from the very beginning, over a period of at least 3.5 billion years, has shared the same four nucleotides making up the DNA molecule. It is a truly amazing story – a *true* story, based on a huge trove of evidence developed by our amazing reasoning brain. And it is and will remain true for all time, whether or not anyone alive on earth accepts it as such.

To be sure, this great adventure began with the ancient Greeks: In Plato's *Theaetetus* Socrates notes, "The only beginning in philosophy is wonder," and in opening of his *Metaphysics* Aristotle remarked, "All men by nature desire to know." But for many centuries thereafter the remarkable Greek heritage in mathematics and the natural sciences languished, except in some of the early Islamic societies, until it began to revive again during the Italian Renaissance, a "rebirth," in the fifteenth century, where the great figure is Leonardo da Vinci, and where this adventure was reborn, unsurprisingly and quite literally, with new translations of many of the ancient Greek scientific and mathematical manuscripts.

Three more centuries passed before leading figures were able to state clearly what was distinctively "new" about the reborn mathematics and natural sciences; here the classic works are Francis Bacon's *Novum Organum Scientiarum* [*New Instrument of Science*] (1620), Galileo's *Discourse on Mathematical Demonstrations relating to Two New Sciences* (1638), and Giambattista Vico's *The New Science* (1725). Another two more centuries of intense intellectual labor were needed before it could be said with assurance that there would no longer be any viable competition for modern science, whether from theology or any other source, as an explanation of the workings of the natural world.

It goes without saying that this science will be incomplete to some extent, quite possibly indefinitely into the future. But it is also the case that, whatever further theories and experimental findings are established, they will result from exactly the same (gradually refined) methods that produced the earlier ones. One of the greatest accomplishments of the new science was its steady incrementalism, over centuries and generations of practitioners, wherein earlier syntheses are incorporated into newer and wider ones – an ongoing development not necessarily in a straight line, but rather often with some twists and turns. Newton's cosmology was not cancelled by Einstein's, but rather incorporated as a special case within the later one, just as, one day, Einstein's will be subordinated, but not cancelled, within a later and wider "theory of everything."

For a creature that has been gifted by nature with a thinking brain such as ours, *this* knowledge – the knowledge about the evolution of the universe over almost 14 billion years, and the evolution of *homo sapiens* over the past 3.5 (or so) of those 14 billion years – is precious beyond all reckoning. (This is most certainly not a claim that this entire evolutionary story necessarily led to the end-point of human intelligence, or that our intelligence is somehow worthier than any other end-point. We are fated to disappear once and for all, along with our lovely planet, sometime in the future, as our knowledge of astrophysics assures us.) If we count the time from the ancient Greek thinkers down to our own age, attainment of this knowledge required more than twenty-five centuries of hard intellectual effort by tens of thousands of individuals.

It is a treasure to be preserved at all costs, for there is no certainty that, once lost, it could ever be recreated. It stands alongside the other treasures we have inherited, our fine arts, our history, our architectural wonders, and the modern technologies that make our lives comfortable and our minds able to enjoy all these treasures. We have a sacred duty to preserve and protect them for the enjoyment and enlightenment of all future generations.

To reiterate the thesis: Modern science and its evidence-based methodologies represent a permanent, unambiguous good for humanity. Not so with its closely-allied forces, technology and industrialization. This distinction may be puzzling to many, for

the simple reason that the three allied forces seem to be so deeply, and seemingly irrevocably, interconnected as to rise or fall in tandem. And, in a sense, that is precisely the problem! When a scientific discovery is married to an emergent technology which together promise great new benefits to humankind, and then this duo is successfully scaled-up so as to be mass-produced at a price affordable to many people – as has been demonstrated countless times – why would the overall end result not share the status of an unambiguous good with the initial discovery itself?

The answer is, once a wide-ranging discovery/technology enters the social world as one or a series of commercial products that better satisfy human needs and wants, its character inevitably changes. It is no longer a simple increase in the accumulated human understanding of how nature works; rather, it has become an intervention in the social world *whose wider effects cannot easily be predicted or controlled*. To take a few well-known examples: Zyklon A was invented in Germany in the early nineteen-twenties as a cyanide-based pesticide for insect control on the basis of general scientific discoveries in chemistry as well as technologies that would make its use cost-effective in commercial quantities. A later product, Zyklon B, was a similarly cost-effective product when used to kill millions of prisoners in the Nazi extermination camps, an application not remotely foreseen by the developers of the earlier one.

The German chemist Fritz Haber, whose lab devised Zyklon A, earlier won a Nobel Prize for his great breakthrough in fixing nitrogen from air to create ammonia; his colleague, Carl Bosch, worked out the method to scale up its production and mass-produce synthetic fertilizer. This application vastly increased the capacity of agriculture to produce food, resulting in huge increases in the human population. But ammonia also was used to make high-explosive material for shells and bombs, vastly increasing the death tolls in World Wars I and II. Or, one later example: The science and technology of gene therapy is approaching the point where medical interventions during embryo development could eliminate entirely, or reduce the frequency of, the occurrence of certain inherited diseases that have truly devastating impacts on human life. As soon as this prospect was on the horizon, other calls were heard to allow gene

enhancement by the same means, whereby those who could pay for the treatment might receive significant advantages over untreated individuals in the competition for wealth and success in society.

Similar examples could be multiplied endlessly. But perhaps it would be better to step back from specific examples and look at the broader picture. At the level of pure scientific discovery, atomic and subatomic physics was recognized as a marvel of conceptual and applied human intelligence by the end of the 1930s. But by the middle of the twentieth century, the science-technology-industry-economy-military nexus related to atomic physics had arrived at the point where it was possible to imagine two utterly opposed future scenarios simultaneously: one, where the technology of nuclear fission would produce electricity for industrial and home uses that would be "too cheap to measure," ushering in period of boundless human prosperity the world over; the other, where nuclear war promised the complete destruction of all advanced societies – of human civilization "as we knew it" – a function of that larger nexus that was, and to some extent remains, a looming disaster, a truly catastrophic risk scenario.

Some may take comfort in the argument that, since the worst-case scenario never happened, we can put the whole business out of mind. Yet there were some very near misses in those days, when the threat suddenly came closer to realization, and not only during the Cuban Missile Crisis of 1962. And in fact, although most people would not have realized it, the risk of nuclear war between these two superpowers actually increased after the Cold War ended, because one of them, Russia, had been significantly weakened militarily, in terms of its overall capabilities, after the disintegration of the Soviet Union, leaving it more heavily reliant on its nuclear arsenal if a new world war had erupted.

Technology and industry, as opposed to the pure sciences, represent in their very nature a double-edged sword. This much is not new, of course, and one often hears the rejoinder that goes something like, "Well, we have to encourage the good outcomes and prohibit or control the bad ones." Indeed. In the bigger picture, however, two evident facts expose the silliness of that rejoinder. First, the overall potency of modern

human technologies gradually was enlarged to the point where failing to control the downside had potentially adverse consequences, in terms of death and destruction, on a gigantic scale. Second, efforts to control the downside risk by formulating enforceable international agreements simply could not keep pace with the escalating threats. Some notable successes, for example the convention on biological and chemical warfare, were more than offset by other notable failures, particularly in the areas of climate change and nuclear disarmament. Carrying out widespread genetic manipulation of physical and especially brain function, done in the germline so that one-time alterations would be heritable indefinitely into the future, was a certainty once it had become reasonably reliable in achieving the intended effects.

And then there was perhaps the most bizarre innovation program of all, namely the wish to create a "super-intelligent" entity, the sheer thinking power of which would dwarf the abilities of ordinary human brains, including the most creative intellects known to us: One overenthusiastic proponent referred to the possibility of creating endless numbers of "Einsteins and Beethovens"! Quite apart from the inherent vulgarity of the idea of manufacturing such figures by the dozens, the author overlooked the complementary possibility of our instead encountering endless numbers of Hitlers and Tamerlanes. There were various tracks laid out for realizing this objective, one of which (genetic enhancement of brain functions) has already been mentioned. But this was the least-favored option, because of the length of time needed to carry out the manipulations over a number of human generations. So, the fondest hopes of its champions were placed in two other initiatives, first, perfecting "whole-brain emulation," whereby a living brain would be digitally imaged and scanned, whose functions could then be manipulated at will to increase its capacities.

And second, forging a purely mechanical construct combining steady advances in computer processing speed and power, artificial intelligence or AI (mimicking the brain's neural nets), and software-based emulation of the neural processes that occur in the brain's limbic system. (The limbic system – including the hippocampus, thalamus, amygdala, cingulate gyrus, and other regions – is the interface between the brain's

subcortical and cortical regions and regulates important functions such as emotion, behavior, memory, and decision-making.) Then, having created super-intelligent entities of this kind, which had integrated their mastery of the full range of human emotional states with their vastly superior data-processing skills, the remaining pathetic, backward, and under-powered representatives of *homo sapiens* would see clearly the benefits of being "merged" with their benevolent mechanical *doppelgängers,* becoming human-machine chimeras – surpassing the ancient human fantasies about animal-human hybrids – and live happily ever after. "We will merge with our technology, … [the] future superintelligent A.I.s," promised Ray Kurzweil, one of the chief dreamers. Perhaps one might have even labeled the new species *pan*, after Pan, the goat-man, the Greek god of the wild realm, were it not already reserved for the name of our nearest cousins, the chimpanzees (*pan troglodytes*) and bonobos (*pan paniscus*).

The partisans of this beautiful idea enjoyed pointing out the superiority of AI in terms of simple information-processing speed, with electronic signals traveling at just below the speed of light (about 300 million mps, or meters per second), whereas our pathetic brains can only manage about 120mps. Among the partisans who were afraid of missing its coming into being during their own allotted time on earth, the fond hope was that the evolution of these mechanical entities would arrive at a point of recursive exponential growth, completing their own triumph much sooner than otherwise expected. The most astonishing claims made for this constructed entity were that it might develop an autonomous will, uncontrolled and uncontrollable by human agents, and moreover that it would be clever enough to conceal the growing capacity of its own will, deceiving the humans who had originally created it – including an ability to infiltrate networked electrical power delivery systems and network-control mechanisms – until it was no longer possible for any human agent to use a "kill switch" to shut it off or seek to destroy it.

I will have more to say about the quest for superintelligence later, in Section Two. For now, I just wish to focus on the curious way in which some of the proponents of this

initiative presented it to the wider public. In essence they said, "It's coming, get ready, it will be great, and by the way, resistance is futile, you couldn't stop it even if you tried." They were appealing to an old idea, sometimes called "the technological imperative," which held that advances in mechanical systems, once painfully slow in human societies, and then steadily accelerating in the science-technology-industry-economy-military nexus, were unstoppable so long as that larger nexus remained robust.

Its proponents readily conceded the point that all technologies had the capacity to result in bad as well as good outcomes, and that society would be challenged to find the ways of preventing or minimizing the bad (the downside) while reaping the benefits of the good (the upside). But what they would not concede was that it might be quite appropriate to say, "No, we should just stop here, we have enough in terms of a capacity to provide a reasonable level of goods and services for a satisfying human life." The proponents refused to recognize the argument that technologies had arrived which would have catastrophic outcomes if the downside risks simply could not be managed – or that humanity might only discover that the international community was unable to manage the most serious downside risks, those that threatened the future of civilization itself, until it was too late to avoid them.

The idea of the "technological imperative" is an old idea, originating long ago in the mind of the philosopher Francis Bacon, who understood early on that if the "new instrument of science" were to win the day (which it did, eventually), it would create a desire for "the effecting of all things possible," that is, an endless growth in the capacity of human society to manipulate and control natural processes. Just before his death he wrote a charming utopian fantasy, *New Atlantis,* first published in 1627, in which he imagined a society that was ruled by a scientific and technological élite, who were unopposed by their fellow-citizens because of the continuous stream of benefits derived from their innovations. He inspired a long tradition of thinkers for whom any attempt to dam this stream was regarded as morally wrong or even perverse. If as a result certain

types of problems (adverse outcomes) arose, as they inevitably would, they could be fixed – usually by devising even newer technological solutions.

And throughout the history of modern industrialized society such problems often were fixed by means of new technologies, as, say, less polluting solutions were found for energy and goods production. But as the technologies became intrinsically more powerful, entailing severe catastrophic risks as unintended by-products, the solutions proved inadequate to the task (as in the cases of climate change, nuclear weapons, and genetic manipulation). Thus, there arose a kind of "terror of the technological imperative," a sense that humanity had become locked into a path of development that was out of control, one where dominant social and economic interests steadily raised the ante in the game of seeking enhanced powers of manipulation over natural processes as a solution to intractable issues raised by earlier bets placed in that game. At some point the music accompanying this merry-go-round was bound to stop, and it did, when climate-change denial finally reaped the whirlwind, and billions of impoverished people fleeing the rising seas and ever-more-destructive storms brought industrial society – and its unshakeable belief in the mantra that every problem had a technological solution – crashing down in much of the world.

In the end, the idea of a technological imperative was exposed as being nothing more than an article of faith, a secular religion, despite its veneer of scientism. Like the monotheistic creeds, it told of an inescapable fate for humanity; no questions were allowed, only submission; it was, in other words, a form of voluntary servitude. Great benefits were assumed, the risks largely ignored, the net-benefit calculus obscure. In its historical evolution, this imperative (famously called the "iron cage" of rational action by Max Weber) had been born in a bitter struggle against an archaic concept of nature as seen through the religious lens of the Abrahamic creation story. This is why, in one of the most famous episodes in the emergence of the "new science," the Catholic Church had come to regard Galileo's defense of the heliocentric model of our solar system as a grave threat to its faith-based worldview, wherein the earth that God had created for

humankind necessarily must be at the "center" of everything, both cosmologically and metaphorically.

Ultimately the new science vanquished its opponent, asking only, without presupposition: How does nature *work*? It had sought thereby to simply bypass the religious idea that there was an intrinsic *moral* meaning in the conception of how nature operates. But in the process of doing so, it had fought so long and so strenuously to exclude any consideration of intrinsic value in its operational concept of nature, that it was left without any ethical anchor for itself – that is, any way to judge the worth of yet another increment in instrumental power as measured against some notion of a truly "human" life.

Once the old value-laden shackles had been cast off, there was, it was felt, no need to forge new ones. After we have "merged" our brain with the artificial superintelligence apparatus, there will be – apparently – no limit to our computational wizardry. Is it just silly to inquire: Will we be *happy* in this new state of being? Or will we all turn out to be living with an extreme, permanent form of bipolar disorder, with our attached signal-processing units, operating near the speed of light, impatiently, maniacally overriding our antiquated neural circuitry in order to handle the incoming stream of data inputs more efficiently? What does *happiness* have to do with anything, anyway? On the other hand, forget happiness, that may be asking for far too much; how about just a little peaceful *sleep*, before full-blown psychosis sets in?

Already in the first half of the nineteenth century, the end-result of the future trajectory implied by the new science-driven technologies, made possible by steam-power, was clear to many: the complete mechanization of labor, the coming of the "age of machinery," whereby humans would be reduced to nothing more than the servants – or indeed, the slaves – of their own creations. *This* was perceived as a great rupture in human affairs, ominous and terrifying in its implications. (*The reader will discover in the next chapter an essay on this topic, entitled "Sublime Machine," written almost a century ago but still worth reading*.) The unsettling implications in the age of machinery were assimilated into the transition from utopian to dystopian literature at the outset of

the twentieth century with the publication of the great short story by E. M. Forster, "The Machine Stops," in 1909.

By the early part of the twenty-first century it had become crystal-clear, to all those who had eyes to see, that there never was, and would never be, any *unambiguous* good for humanity in the advanced technologies and mechanization laid at its feet by the modern natural sciences. By no means does this suggest that no good at all is to be found there! There are, first, the simply incalculable benefits we derive from an abundant supply of four necessities: safe water, electricity, heat, and air conditioning. Then there is the relief from suffering through medicine and dentistry; the mental, physical and longevity advantages of increased nutrition and the control of infectious diseases; the possibility of an end to onerous, backbreaking labor, as well as child labor; sufficient leisure for contemplation and education; safety and security of the person, especially for women: All these supremely important benefits, and many more, potentially available to everyone, everywhere on the planet, cannot be had without modern technologies, adequate energy resources, and plentiful machines.

Many of the most important benefits had begun to be widely available in economically-advanced nations by the turn of the twentieth century, and others were already on the horizon. But soon the turn came, and the remainder of that century combined the most destructive wars in human history along with, in their aftermath, the threat of nuclear annihilation. The total destructive power of the competing nuclear-missile arsenals was sufficient to obliterate all modern cities many times over, to deposit enough long-lived radioactivity to make their locales permanently uninhabitable thereafter, and to bring about a "nuclear winter," spreading widespread misery around the globe perhaps for centuries to follow. No one ever explained what was the point of this massive overkill capacity.

But it is a peculiarity of the human mind that threats narrowly averted, however dire, barely register in the thinking about future risks. Many experts caution us against excessive risk-aversion, but few seem averse to reckless risk-taking, such as in the run-up to the global financial crisis in 2008, the actual adverse effects of which were bad

enough, but also another case where far worse downside scenarios were just barely avoided.

Well before the first half of the twenty-first century was "in the books," so to speak, it had become obvious that this recklessness had to stop. It was obvious that, taken as a whole, the collective authority of the world's nation-states was unable to safely manage the prevailing risk-risk and risk-benefit trade-offs spawned by the fecund nexus of science, technology, industry, economy, and the military. The clearest sign that a tipping-point had been reached, a point where the idea of the technological imperative finally revealed its profoundly irrational and self-destructive core, was to be found in the delusional hopes placed in superintelligence and gene enhancement.

For the fans of these solutions had openly embraced a future for humanity in which what was "human" – as examined and articulated across 2,500 years since the ancient Greeks – was to be casually and unceremoniously dumped into the trash-can of history. And there would be no going back, no occasion for remorse and reconsideration, once the experiment had been launched, at least not without the slim chance that, *homo sapiens* having self-destructed, we would be allowed by Mother Nature to try again, with a new branching off the old hominid line, as had happened five million or so years ago, a branching off that evolutionary line where those clever bonobos, *pan paniscus*, reside. Alas, there was a greater likelihood that we would have already wiped out the bonobos once and for all.

These two looming catastrophes arising out of a technological *hubris*, a colossal failure of imagination in which an expiry date had been affixed to the human essence, or species-being (*Gattungswesen*, to recall Marx's nice formulation), apparently had become inevitable because a final solution to past mistakes in submitting to the technological imperative could be found only in making yet another and far more serious one. Some of us concluded that this "solution" had to be averted at all costs. So, responding to the collapse of major institutions in industrial societies, a handful of small like-minded groups resolved to strike out on an entirely new path, as described in the two earlier volumes of this trilogy. That route will be laid out systematically in

Section Two, “Pathways to Utopia”; here I wish only to complete my brief account of the historical rupture that took place in the modern West, contrasting the true one (modern science itself) with its misleading counterpart (the science/technology/industry/economy/military nexus).

Science appears in both perceived forms of the rupture because it has carried a dual meaning for its champions ever since its seventeenth-century beginnings. The first was, simply put, an insight into a method which could unveil the inner workings of natural processes – not at once, immediately, in a single flash of comprehension, as the alchemists had hoped – but over time, laboriously, patiently, collectively, incrementally, slowly sifting confirmable evidence from mere speculation, slowly building up a “weight of evidence” rather than cherry-picking pieces of evidence to advance a preselected theory. Francis Bacon, as usual, gave us a whimsical but accurate simplification of this method: Just follow nature by careful observation and experimental trials, focusing on what is repeatable and ignoring extrinsic details, until you have seen how specific initial conditions lead to specific results or end-points via specific processes, and then you (or anyone else) can reproduce the results – say, the discovery of the elements hydrogen and oxygen in the eighteenth century – at will. And then, having found these very intriguing substances, you can put them to use in the service of human needs.

The second perception of science’s meaning arose simultaneously with the first, and Bacon was the primary author of this one too. It held that this scientific method would bestow on humanity “power over nature.” This was always a curious formulation and at first glance makes little sense: How does following and reproducing the results of natural processes grant us power over them? The solution to the apparent conundrum is that by being able to reproduce a desired result, for example, a useful chemical reaction, at will, we have increased human operational capacity in the world, that is, the ability to turn knowledge into the power to make new things and new ways of making things (technologies). This represented a power – actually, in cooperation with, not over, nature – to enlarge human agency and therewith human resources, desires, and populations. Ultimately, it was hoped, we could find some way to do anything we

wished to do, from flying, eating more meat, or more efficiently slaughtering other people, to teleportation, time travel, waging intergalactic warfare, freezing our brains for later revival, or living forever.

Not everything anyone wishes to do is *ipso facto* a worthy objective, obviously, so long as we are willing to apply some value-framework, however minimal – say, the Golden Rule – before carrying it out. In other words, not every wish to exercise a power we are able to deploy in the world is intrinsically worthy. And that is in my estimation the simple difference between science as the true understanding or nature and science as the enabler of new technologies of power in the world. The first is and always will be intrinsically good; the second cannot ever be (although any particular application of it *may* be so, if it occurs in an appropriate values-context).

The first is both a collective and an individual good. It is a collective good because, as the Enlightenment thinkers argued, it mitigates the unfortunate propensities of humans to torment each other on the basis of superstitions. It can be an individual good for at least some individuals, for it bolsters the pattern-seeking proclivities in our cerebral cortex by uncovering the regularities and causal structures hidden behind the world of appearances. For those individuals who cannot accept it because it conflicts with the story of faith, they are entitled to go their own way, so long as neither group seeks to impose its convictions on the other.

It should occasion no surprise that, some five hundred years after the new science first took hold in Western Europe, its adherents are far outnumbered by those of the religious communities (although, as has been noted, there are some who live by both). For this new science represents a truly radical rupture with the understanding of nature that preceded it and that had flourished in different forms for many millennia since permanent human settlements were first established. In the early years of its development it spread its tentacles through its social host so unobtrusively that dominant institutions were slow to realize what was happening; by the time that its subversive method openly launched challenges to long-established beliefs, in the second half of the nineteenth century, this genie (as an integral part of a larger nexus),

was delivering far too many concrete benefits for anyone to seriously suggest that it be put back in the box again – at least not until, during the first half of the twentieth century, a powerful and horrifying reaction to Enlightenment philosophy arose, a story that is told in a later chapter.

The philosopher Hegel coined a powerful metaphor to explain this subtle, subterranean infiltration of reason in history which, he said, operates "behind the backs" of individuals and societies. What he meant is that the true significance of important historical transitions is not grasped until long after they have established themselves. The rupture represented by modern science is a classic instance of this type.

But it represents something else, too, in which it is unique: The mathematics and geometry of ancient Greece was the first true universal in human thought, and when modern science revealed its dependence on mathematics – for the first time, in a systematic way, in Galileo's summation of his life's work, *The Discourses and Mathematical Demonstrations Relating to Two New Sciences* (1638), it became the second. On the other hand, despite their pretentions to universality, major world-religions, especially the proselytizing ones such as Christianity and Islam, never succeeded in conquering the globe, and almost certainly never will. But anyone, anywhere, at any time, who wishes to grasp how nature works, on the planet earth as well as in the vastness of time and space across the universe, must use the method of the new science – a method still evolving, to be sure, but one that has always built incrementally on its earlier stages in order to advance.

But what happened to the original Enlightenment promise of the broad increase in *public* understanding that would result from the deployment of the new instrument of science? The answer to that question is: twentieth-century physics (as we shall see in chapter 3).

Chapter 2: Sublime Machine

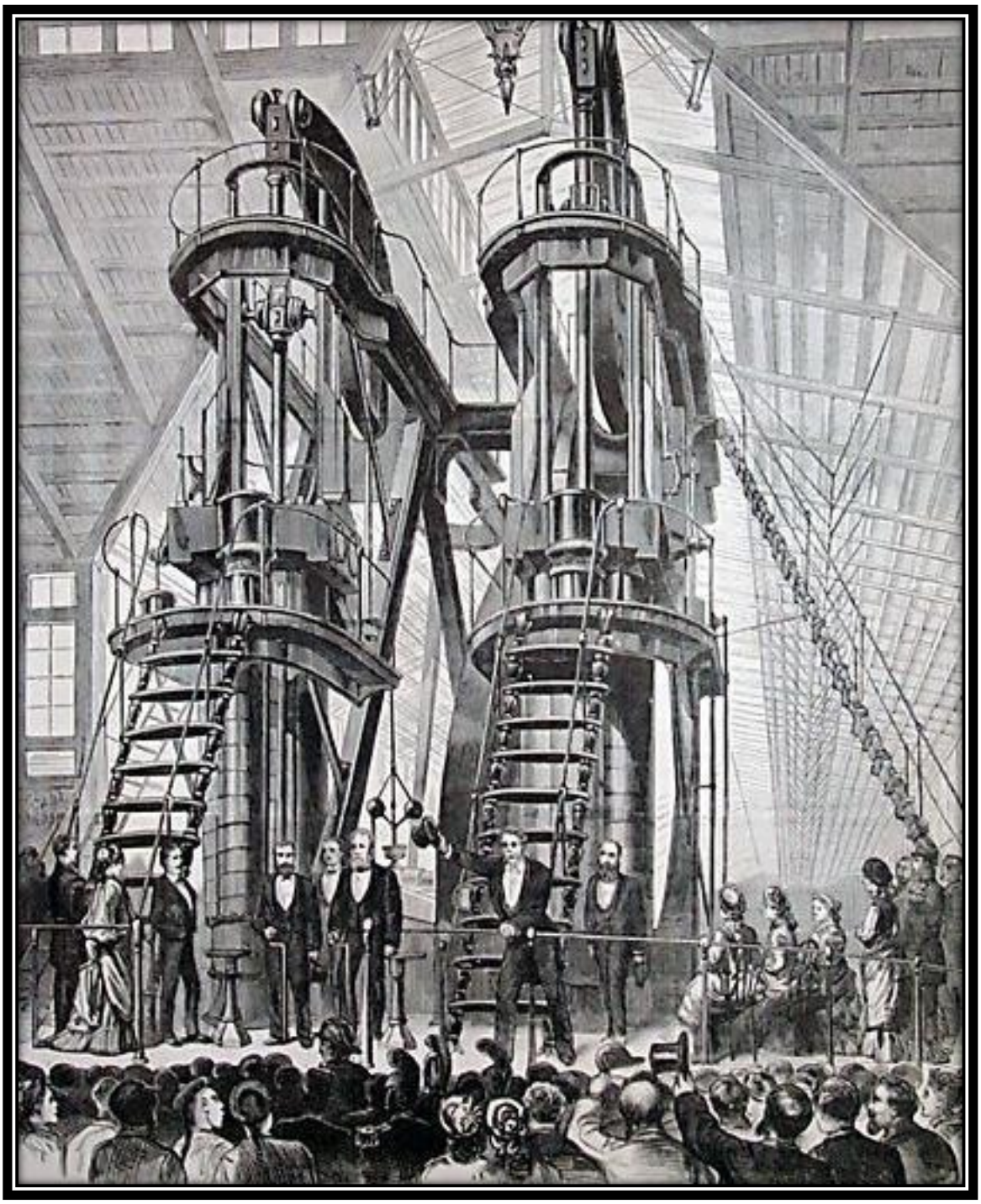

Figure 4 President Grant & Emperor of Brazil, at the Corliss Steam Engine, Philadelphia Exposition, 1876

THE SHOW-STEALING EXHIBIT at the Philadelphia Centennial Exposition in 1876 was a Corliss steam engine, weighing 680 tons and standing thirty-nine feet high, which provided all of the power for the entries in Machinery Hall. According to contemporary accounts, its presence overwhelmed all who entered the hall, whether they were ordinary fair-goers, such high and mighty as President Ulysses S. Grant and the Emperor of Brazil, or well-known writers like William Dean Howells. It excited the popular imagination, as had other such events beginning with the Great Exhibition in 1851, and so outstripped the capacity of ordinary descriptive reporting that only ecstatic metaphorical construction could register reactions to it. John F. Kasson notes that the fair-goers' descriptions of their experience "frequently became incipient narratives in which, like some mythological creature, the Corliss engine was endowed with life and all its movements construed as gestures. The machine emerged as a kind of fabulous automaton – part animal, part machine, part god."

One guidebook for the Philadelphia exposition offered its readers a lesson in aesthetic judgment. Whereas traditionally poets located the experience of the sublime in our reactions to wild nature or powerful human passions, the guidebook claimed that the modern age recognized the sublime in the design and operation of its great machines. And a newspaper reported that in the presence of the Corliss engine "strong men were moved to tears of joy."

Almost exactly one hundred years later the French "neo-Dadaist" artist Jean Tinguely persuaded the director of New York's Museum of Modern Art to offer the museum's sculpture garden as the site for a spectacular auto-da-fé by Tinguely's self-destroying machine. (The performance was named *Homage to New York*.) When finished, the machine was twenty-three feet long and twenty-seven feet high; its main distinguishable components were a piano, an old Addressograph machine, eighty bicycle wheels, steel tubing, a meteorological balloon, a huge klaxon on wheels, a wide assortment of small mechanized devices, and various chemicals – smoke, flash powders, and foul-smelling substances.

When the main motor was switched on, the piano keys were struck, wheels turned, klaxons sounded, a radio blared, clouds of smoke billowed forth; a number of small constructions broke free and wheeled about; and small objects were hurled through the air. Then the piano caught fire, the steel tubing supports began to give way, and the terrified museum authorities ordered in firemen with axes and extinguishers to finish off the machine. Once set in motion, the machine's self-destructive orgy had followed pretty much its own course, rather than the artist's specific sequence of events, and this spontaneity was precisely what Tinguely had hoped most to achieve. To him this machine "was the opposite of the skyscrapers, the opposite of the Pyramids, the opposite of the fixed, petrified work of art, and thus the best solution he had yet found to the problem of making something that would be as free, as ephemeral, and as vulnerable as life itself." The late machine was described as both a beautiful and a terrible thing, and it was reported that at the end some spectators had wept.

All in all, the concept of the sublime – the ineffable union of awe and dread, terror and attraction – is a good a guide as any to unravelling the modern reaction to industrial society and the machine. The iconography of the machine supports the case. Kasson remarks that many nineteenth-century popular illustrated magazines chose a graphic style and accompanying text for their drawings of large machinery that heightened the sense of "mystery and majesty." One of the most famous illustrations was J. O. Davidson's *Interior of a Southern Cotton Press at Night* (1883). Davidson himself supplied the following explanatory note: "Beneath the converging rays of electronic lamps and reflectors a most weird effect is produced, for the machine assumes the aspect of a grand and solemn demon face, strangely human, recalling the famed genii of the Arabian Nights." In the great scene in Fritz Lang's film *Metropolis* (1926), where tier upon tier of identical machines, deep underground, are attended by workers whose rhythmic movements follow those of the levers and dials, the machine's face closely resembles the one engraved by Davidson.

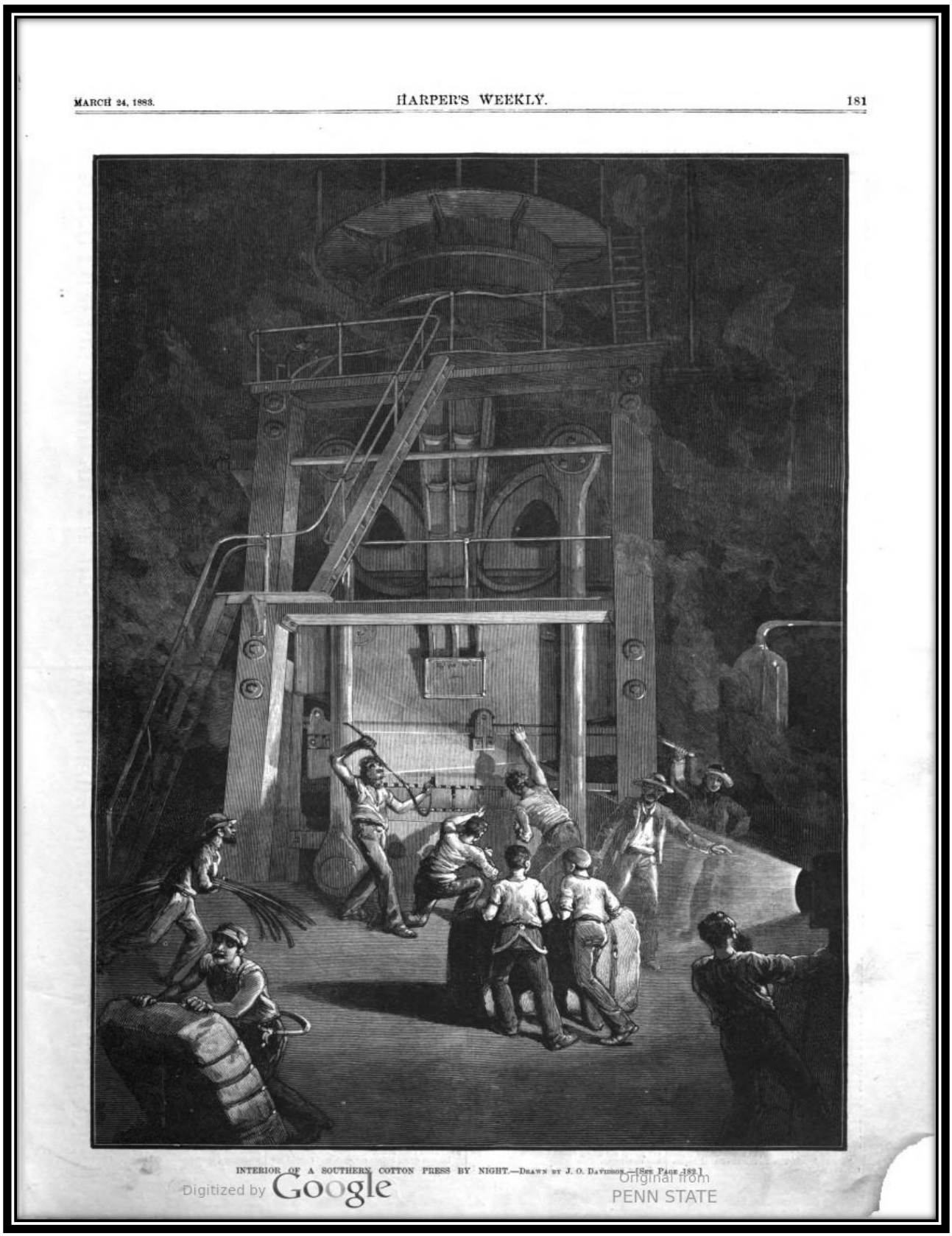

Figure 5 J. O. Davidson, "Interior of a Southern Cotton Press," Harper's Weekly, 24 March 1883

The iconic representation of the machine, in eliciting the feeling of the sublime, testified to the darker side of the human experience with large-scale machinery that qualified the popular enthusiasms expressed at the great exhibitions. This popular ambivalence was mirrored in the struggles by imaginative writers and social thinkers to come to terms with the industrial age.

The majority of nineteenth-century political economists, and virtually all the marginalist economists who created a formalized discipline after them, typified the "happy consciousness" of industrial society: they were satisfied that the abundant and manifest benefits supplied by industrialism and the division of labor overawed whatever negative aspects inevitably accompanied them. They never entirely silenced the dissenting voices, however, who worried about the moral degeneration and degradation of skills in the labor force. Originating in a striking passage in Adam Smith's *Wealth of Nations* (1776), this dissenting strain was kept alive mainly in the nineteenth-century socialist movement, notably by Robert Owen, Karl Marx, and William Morris. It remains alive in the twentieth-century tradition that runs from Thorstein Veblen to Ivan Illich.

Many dissenting social thinkers believed, however, that the degeneration characteristic of industrial society was remediable, in most cases by more or less drastic reordering of economic and political circumstances. It was much different with those who represented the predominant aesthetic sensibility of their time, for among them the prevalent mood ranged from dismay to horror. Beginning about 1830, when the impact of industrialism began to register, major writers entered the lists against the machine and the industrial age. Thoreau, the later Emerson, Melville, and Henry Adams in the United States; Zola, Balzac, and Flaubert in France; Heine, Hesse, and Thomas Mann in Germany; in nineteenth-century England, Carlyle, Dickens, Ruskin, and Morris, and in the early twentieth century Forster, Lawrence, and Huxley. For some of these it is (at least overtly) a minor theme, but for others the machine becomes the symbol of

degeneracy itself. This mood's culminating expression is the great anti-utopian novel of the early twentieth century, Yevgeny Zamyatin's *We*.

The anti-industrial sentiment also predominated in major English and European developments in the plastic and decorative arts, in part as a reaction against the influence of industrial design on public works and consumer goods. The Aesthetic Movement and Art Nouveau set their faces resolutely against mechanical reproduction and industrial design. Only in the 1920s did architecture and design begin to reconcile themselves to the industrial age.

One can date the aesthetic reaction to the machine from 1829, when Thomas Carlyle's great essay "Signs of the Times" baptized his period the "Age of Machinery." This reaction is completed almost exactly a century later, with the publication of the two greatest anti-utopian novels, *We* (written in Russian in 1920, but published first in English translation in 1923); and *Brave New World* (1932). George Orwell was the first to identify Zamyatin's theme: "What Zamyatin seems to be aiming at is not any particular country but the implied aims of industrial civilization…. It is in effect a study of the Machine, the genie that man has thoughtlessly let out of its bottle and cannot put back again." An allusion to the genie, which we have already encountered in J. O. Davidson's commentary on his illustration of the cotton press, is itself one of the most common textual threads in the literary response to the machine age.

The aesthetic response to industrialism after 1830 argued the shallowness of other reactions, especially in political economy and social thought. The latter were, as suggested above, divided into a predominant "happy consciousness," which welcomed industrialization with open arms, and a dissenting minority, which wanted urgent institutional changes to counteract its deleterious impact on labor and social relations. Most of those in the latter category, however, contended that these negative aspects could be overcome and that the machine age could be turned unambiguously to mankind's benefit.

The dominant literary metaphors appeared to rule out this eventuality. For at its deepest level the matter appeared to be one of life and death, considered in terms of

the essential determinants of what it means to be human, and the machine seemed to represent the ultimate degeneration, the death of humanity. In the following pages this theme will be tracked through a series of metaphorical constructions that lead inexorably to the opposition of life and death.

Root Metaphors

Sander L. Gilman has used the idea of root-metaphor as a way of understanding both continuities and variations over time in literary expressions that reflect common experience. It seems that we often require a means of synthesizing our perceptions of complex events, especially when we are faced with startling new circumstances that fall outside the realm of our ordinary experience. Metaphors – "it was like a thunderclap" – allow us to capture a novel or extraordinary event in forms of thought that are well known to us, thus "domesticating" it; furthermore, they encourage us to believe that we may communicate our experience to others. There is a concomitant risk, of course: metaphorical constructs limit our ability to assimilate new information, because we try to squeeze the unusual into familiar and comfortable form.

That established ways of life are challenged by unremitting technological novelty is something of a cliché by now. Yet we who have become so jaded should not forget how profoundly unsettling was the sprouting of large-scale machinery and the factory system for both society and culture in the nineteenth century. For most people, common folk and artists alike, it was as the world itself had come unhinged. Many found that they could comprehend its significance only by resorting to metaphorical expressions rooted in thoroughly familiar structures of experience. When one recalls the enormity of the changes wreaked in the social and physical landscape in such a relatively short time, it is unsurprising that the search for adequate expressive modes should terminate in the fundament itself: life against death.

No simple scheme can hope to capture all of the varieties of expression for such a universally felt experience. The one to be explored here seems to catch a sample of reasonable size and quality, although undoubtedly much that is equally important slips

through its mesh. The scheme is composed of three levels of metaphorical construction, internally related to one another, which proceed from the "surface" realm of familiar social experience to the ultimate duality of life and death.

The root metaphor for the surface level of representation of the relationship between humanity and the machine is *master and servant*. This had two quite obvious advantages. First, it was a relation that was thoroughly familiar in social experience everywhere. Second, and perhaps more important, it is a relation that is readily reversible in imagination. The affirmative response to industrialism trumpeted the machine as the perfect servant of human objectives, as the long-sought deliverance from necessity and want. The rejoinder quickly made itself heard: the servant will be master. The imagery of the "sorcerer's apprentice," together with that of the Arabian Nights and its genii, have been favored to make the point.

The root metaphor for the second level is further development of the master-and-servant theme. Domination and servitude are external relations in which each side is necessarily the opposite of the other. On the second level, we pass beyond the purely external relation, and the two participants in the human-machine relation begin to switch roles: human agents adopt more passive roles in step with the growing virtuosity of machinery. Machinery based on advanced designs is capable of self-regulation and self-adjustment; at the same time, the human agents who tend the machines have less and less to do. There arises the twin prospect of the autonomy of the machine and people as automatons. The second level of representation is therefore *autonomy/ automaton*.

The "autonomous technology" theme is an old and persistent one in Western thought. Conceiving the machine as autonomous is an extension of the master-and-servant metaphor. Here the machine's role in the relation is reversed – servant is now master – in the sense that we have allowed ourselves to become utterly dependent on its productive power in providing desired goods; strictly speaking, then, this is a case of voluntary servitude. In other words, we set in motion a course of events that resulted at some point in our losing control over what we have created: we can no longer

"freely" choose to have it or not. Since we cannot even conceive of doing without its benefits anymore, we are beholden to our apparatus, and we begin to adjust our behavior to its modus operandi. In Carlyle's words: "Men are grown mechanical in head and in heart, as well as in hand."

What began as an external relation is now an internalized process, whereby the dependent member (the human being) surrenders its own authentic being to its erstwhile instrument. The relation itself and the tension between its originally opposed sides dissolve as society begins to mimic the way machines operate. In *We* Zamyatin gave the most striking representation to the process of internalization and the root metaphor of autonomy/automaton: society is ordered on the model of the machine, and men and women are its subordinate parts, whose "functions" are determined strictly in relation to their role in the apparatus as a whole.

The third level of root metaphor was a direct outcome of what preceded it: the concept of automaton led directly to the imagery of the opposition between life and death. This metaphor works on the identification of the machine with inorganic matter, necessity, repetition and identity, and thus death – and the concomitant association of life with organic processes, and with contingency, variation, or freedom. The machine as automaton, however, possesses characteristics both animate and inorganic: in crossing over the two realms it appears to draw what is living inexorably into the province of the inanimate. Powerful representations of this theme appear in the case studies to be presented later: Melville's "The Paradise of Bachelors and the Tartarus of Maids," E. M. Forster's "The Machine Stops," and Zamyatin's *We*.

Machinery

For industrialism's defenders, machinery had lifted a double yoke from humanity's shoulders, namely, subjection to nature's capriciousness as well as to the corrupting influence spread by relations of dependence among people. Technology would overturn humanity's age-old subordination to physical forces and deliver he realm of nature holus-bolus into its hands, to do with as it would. At the same time, material

abundance and mechanical aids would do away with the employment of people in personal service – an especially prominent theme in the United States, where industrialism had been linked to republicanism. Two years after Carlyle's 1829 essay appeared, its message was thoroughly rejected by a writer for the *North American Review*, Timothy Walker, in "Defence of Mechanical Philosophy." Of the blessings of technology, he wrote: "From a ministering servant to matter, mind has become the powerful lord of matter."

This Baconian theme, both widely sown and deeply rooted by mid-century, was so successful in its propagation because it represented the relation between human beings and large-scale machine technology as analogous with the completely familiar routine of masters and servants. Machines would take the place of servants, who are out of place in a democratic regime; not only could it assume many of the burdensome tasks usually imposed on dependent people, and in many cases perform them more efficiently, but it could also be seen to be more fitting in this role. John Ruskin gave a nice explanation for this point. What a master ordinarily requires of his servants, he remarked, is the maximum output for the least pay (that is, the market value of the servant's labor); and, according to the prevailing economic wisdom, this situation will yield the greatest benefits for society as a whole and all its individual members, including the class of servants.

This would be the case, Ruskin objected, "if the servant were an engine of which the motive power was steam, magnetism, gravitation, or any other agent of calculable force." On the contrary, the servant is a human agent whose "motive power is the Soul," and this fact marks an essential difference: "The largest quantity of work will not be done by this curious engine for pay, or under pressure, or by the help of any kind of fuel which may be supplied by the cauldron. It will be done only when the motive force … is brought to its greatest strength by its own proper fuel, namely by the affections."

Ruskin's distinction reinforces the metaphor of the master-servant relation as a way of understanding the machine's significance for human life, for always lurking in this relation is the potential reversibility of its terms. Thus the machine can be seen as

replacing the human agent and as doing the bidding of human masters. But much folklore also tells of the "reversal of fortune" that catapults erstwhile servants into their master's place to lord it over those who formerly had abused them. Just so the machine.

Melville used the notion of a reversal of roles between humanity and machinery in his portrayal of a New England paper mill in his short story, "The Paradise of Bachelors and the Tartarus of Maids" (1855): "Machinery – that vaunted slave of humanity … here stood menially served by human beings, who served mutely and cringingly as the slave serves the Sultan. The girls did not seem so much accessory wheels to the general machinery as mere cogs to the wheels." This was to become a favorite image in the critique of industrial society, especially in utopian literature that argued for a "second reversal," to be achieved by a radical reordering of social relations to re-establish humanity's hegemony over the instruments to which it had become enslaved. In his utopian sketch *A Traveler from Altruria* (1894), William Dean Howells suggested this in a way that reinforced the root metaphor; in his imaginary future society, "the machines that were once the workman's enemies and masters are now their friends and servants."

The resolution proposed in "re-reversal" confines the relation between humanity and machinery to the first level of root metaphors. It finds adequate the representation given by the metaphor: machines should be regarded as our servants. And it identified our problem solely as one of re-establishing our right to occupy the dominant side in this relation. As we shall see, this seemed a rather superficial resolution to those who wished to consider the matter in terms of deeper levels of significance and more profound root metaphors. For the reversal slips too readily over the circumstances that had given rise to the original reversal, that is, the one whereby human agents had become the machine's servants.

The change in Ralph Waldo Emerson's attitude over a period of twenty years offers a clue about these circumstances. He began with robust confidence in the industrial age and its possibilities for improving the human condition: the enormously

influential essay "Nature" (1836) trumpets that nature "is made to serve." Illustrating what Leo Marx calls Emerson's "rhetoric of the technological sublime" is the following 1843 entry from his journal: "Machinery and Transcendentalism agree well." *English Traits* (1856) records a different sentiment, however: "But a man must keep an eye on his servants, if he would not have them rule him … It is found that the machine unmans the user. What he gains in making cloth, he loses in general power … The incessant repetition of the same hand-work dwarfs the man, robs him of his strength, will and versatility, to make a pin-polisher, a buckle-maker, or any other specialty … Then society is admonished of the mischief of the division of labor, and that the best political economy is care and culture of men."

Industrialization

Emerson's mention of pin-polishing stands in ironic contrast to the famous opening chapter of Adam Smith's *Wealth of Nations*, which had heaped praise on the division of labor and had made Smith's own pin-making illustration a legend in the subsequent political economy literature.

Seventeenth-century Europeans were unable to decide whether the barbarous ways of the New World inhabitants were a degenerate form of earlier civilized conditions or simply a case of arrested development. Their successors may not have resolved this point, but they were confident at least that they knew the proximate cause of their misery. According to Adam Smith, the "savage nations of hunters and fishers … are so miserably poor" because their labor productivity is so low, and this in turn results from their ignorance of the benefits bestowed by the division of labor.

Smith also knew how to reckon the price paid for economic progress, however. The mental faculties of everyone in "barbarous societies" reman "acute and comprehensive" and are not "suffered to fall into that drowsy stupidity, which, in a civilized society, seems to benumb the understanding of almost all the inferior ranks of people." The division of labor confines the worker's activities to routine tasks: "The man whose life is spent performing a few simple operations … has no occasion to exert

his understanding … He naturally loses, therefore, the habit of such exertion, and generally becomes as stupid and ignorant as it is possible for a human being to become … His dexterity at his own particular trade seems, in this manner, to be acquired at the expense of his intellectual, social, and martial virtues." Material progress is won at the expense of widespread degeneration in mental faculties and the capacity for exercising good judgment in public and private affairs.

The Tory critique of industrial society inspired by Carlyle made much of this theme, claiming that the proponents of industrialism and economic development regarded the working population as nothing but "animated machines." Their opposition lent voice in the political arena to the widespread anti-machinery sentiment among the working classes in the early phases of the factory system and to the tremendous social disruptions that accompanied it. The Tory critique's force diminished as it became increasingly apparent that the necessary concomitant to its attack on industrialism was preservation of the traditional agrarian economy and social hierarchy. This left sustained opposition effectively in the hands of the radical critics, who also objected to the degradation of labor and skills under industrialism, but who steadfastly maintained that under radically different social arrangements the highest possible degree of application of machinery to production was in the workers' interests.

Among all those who were willing to commit themselves to this course, Marx grasped best its profoundest implications: "In no way does the machine appear as the individual worker's means of labor … Not as with the instrument, which the worker animates and makes into his organ with his skill and strength, and whose handling therefore depends on his virtuosity. Rather, it is the machine which possesses skill and strength in place of the worker, is itself the virtuoso, with a soul of its own in the mechanical laws acting through it … The science which compels the inanimate links of the machinery, by their construction, to act purposefully, as an automaton, does not exist in the worker's consciousness, but rather acts upon him through the machine as an alien power, as the power of the machine itself … The production process has ceased to be a labor process in the sense of a process dominated by labor as its governing unity."

The laborer ceases to be the "chief actor" in the production process and becomes instead only the "watchman and regulator" over it.

The radical tradition split into two quite different currents in response to the growing presence of machinery in production and the consequent deskilling of labor. The most influential current, in which Marx and most modern socialists are found, accepted the declining role of labor and its traditional skills in producing life's necessities and relegated the cultivation of skill and virtuosity to the realm of free time or leisure. A much smaller branch, for which William Morris's utopian tract *News from Nowhere* (1890) is the chief source, drew the opposite conclusion: re-establish skilled craft labor as the cornerstone of social life and limit wants and satisfactions to what such effort can provide with the smallest possible reliance on mechanical assistance.

The factory system methodically undermined labor's autonomy, its very "substance" as an agent in social life, by eliminating society's dependence on the rich panoply of craft skills heretofore distributed among the working classes. The historical residue of those skills is absorbed by the system of machinery, "whose unity" – in Marx's striking formulation – "exists not in the living workers, but rather in the living (active) machinery, which confronts their individual, insignificant doings as a mighty organism." Regarded in this light, it is easy to see why the master-and-servant metaphor, so readily applied to the relation between humanity and machinery, was also so readily reversible. Having appropriated the essential substance of its putative master, the machine was heir to humanity's accumulated, alienated mastery of its environment; what remained for the "stupid and ignorant" mass of deskilled laborers was only numbing exhaustion in the service of the machine's imperious rhythm.

The radical critique maintained that the machine could be remastered and compelled once again to service mankind's purposes. The system of machinery confronts the worker as an automaton or as a "living, mighty organism" at the level of immediate experience; labor is cowed into submission because it appears as if all skill, initiative, and "virtuosity" have passed irrevocably from it to another kind of being. Its apparent otherness and autonomy, however, upon analysis turn out to be just that,

mere appearance. In truth, it is the same substance: Machinery is "objectified labor," the material legacy of past human skill and exertion, misappropriated in the form of privately owned capital. What seemed to be service to the machine was in fact subjection to another human group that had discovered in large-scale machinery a wondrous device for extracting vast wealth from the labor of others. The realization that labor's enemy was not the machine but the capitalist was for the radical critique the "beginning of wisdom" and the first step toward re-establishing labor's autonomy.

An implicit rejoinder to this program was made in the period under review, governed at the second level by the root metaphor of autonomy/automaton: specifically, internalization of the machine principle in humanity's own mode of being. From this perspective, labor's wresting control of the industrial system away from the capitalists would be Pyrrhic, for this would seal the fate of society as a whole, committed irrevocably to mechanistic modes of action. The very moment of its triumph simultaneously would signal labor's final defeat, and its ostensible autonomy would be a sham. Labor and its skills would be no longer the heart of the production process, since it had surrendered that role to machinery; labor – or what was left of it, namely superintendence – would become just a routine social obligation to earn income for consumption.

In accepting machine production as the dominant means for supplying life's necessities, modern society would be forced as well to adopt a mechanistically oriented routine for life in general:

> The relation in which the consumer, the common man, stands to the mechanical routine of life at large is of much the same nature as that in which the modern skilled workman stands to that detail machine process into which he is dovetailed in the industrial system. To take effectual advantage of what is offered as the wheels of routine go round, in the way of work and play, livelihood and recreation, he must know by facile habituation what is going on and how and in what quantities and at what price and where and when, and for the best effect he must adapt his movements with skilled exactitude and a cool mechanical insight to the nicely balanced moving equilibrium of the mechanical processes engaged. To live – not to say at ease – under the

> exigencies of this machine-made routine requires a measure of consistent training in the mechanical apprehension of things.

These comments by Veblen in his *The Instinct of Workmanship* (1914) were not meant to encourage any hope that this "machine-like process of living" could be overturned. The best that one could do was to take it to its logical conclusion by placing engineers instead of businessmen in charge.

When this concern was first raised, almost a century earlier, it was possible to surmise that the swelling tide of mechanization might yet recede again. The great manifesto for those who so believed was Thomas Carlyle's "Signs of the Times" (1829). For Carlyle, the physical instruments rapidly overtaking traditional productive processes were only the visible expressions of a deeper malaise, namely a habit of mind and action he described in precisely the same terms as Veblen would use much later: a pervasive "matter of factness." The machine itself served as a metaphor for "the great art of adapting means to ends … by rule and calculated contrivance."

Carlyle begins his animadversions by referring to the transformations in the physical environment wrought by the application of machinery to production and transportation. Besides its obvious effects in undermining the craftsman's position, mechanization is faulted for being unable to distinguish between appropriate and trivial applications. By these means people seek to rule nature and in so doing pay a heavy price: "Not the external and physical alone is now managed by machinery, but the internal and spiritual also." Here the machine stands for the disappearance of spontaneity and for the rise of a mode of action that first appraises each situation in strategic terms, then breaks down ultimate objectives into a manageable series of discrete steps, and then assigns means from whatever quarter to the separate tasks: "Has any man, or any society of men, a truth to speak, a piece of spiritual work to do, they can nowise proceed at once and with the mere natural organs, but first call a public meeting, appoint committees, issue prospectives, eat a public dinner; in a word, construct or borrow machinery, wherewith to speak it and do it."

By the time he came to write *English Traits* (1856), Emerson had lost his youthful enthusiasm for the industrial age and was ready to echo Carlyle's sentiments: "Mines, forges, mills, breweries, railroads, steam-pump, steam-plough, drill of regiments, drill of police, rule of court and shop-rule have operated to give a mechanical regularity to all the habit and action of man. A terrible machine has possessed itself of the ground, the air, the men and women, and hardly even thought is free." Taken as a metaphorical allusion, the last sentence could do nicely as an epigraph for E. M. Forster's story "The Machine Stops."

Neither Carlyle nor Emerson, however, was yet prepared to concede that all was lost. There was still time to reverse this disastrous course and reassert the pre-eminence of the natural and the spontaneous over the mechanical mode of action. Despite its deepening penetration of public and private life, mechanization was not yet triumphant over the old ways. Carlyle advertised this hope in an especially revealing way, namely, by suggesting at the end of his essay that the fundamental root metaphor governing the first level of representation was still operative: "Indications we do see in other countries and in our own, signs infinitely cheering to us, that Mechanism is not always to be our hard task-master, but one day to be our pliant, all-ministering servant."

This curious conclusion by Carlyle seriously undermines the force of the argument that preceded it. For it suggests that, however widely it had spread, mechanism had not contaminated the original sources of human action and still could be subordinated to individual and collective ends governed by non-mechanical principles. Or perhaps the opposite is nearer the mark: the force of his own earlier argument undermines Carlyle's conclusion.

Automatons

Matching the uninterrupted march of machine technology in the second half of the nineteenth century was a growing fear that it was indeed out of control. In the relation between humanity and machines, increasingly the former seemed to be the passive partner and the latter the active agent. The more the system of machinery as a whole

assumed labor's erstwhile attributes – skill and indeed virtuosity (Marx) – the more the worker appeared "like a machine" in the derogatory sense, fit only for the dull repetitiveness of routine operations. Emile Zola, who on other occasions rhapsodized about modern technology, filled his Rougon-Macquart novels with allusions to the machine-like and thing-like character of human action and, correspondingly, with the appearance of animate force and autonomous power residing in machinery.

As early as his writings of 1857-8 Marx had referred to an "automatic system of machinery" as the "most complete" and "most adequate" form of the machine itself, "set in motion by an automaton, a moving power that moves itself; this automaton consisting of numerous mechanical and intellectual organs, so that the workers themselves are cast merely as its conscious linkages." The root metaphor of autonomy/automaton, which was to be fleshed out as a favorite device in fiction, alluded not so much to a reversal of roles, as in the case of the master/servant metaphor, as to a complete collapsing of the two sides of a relation into a synthetic entity that transcended both. Its most effective representation was the man-like automaton.

Herman Melville's story "The Bell-Tower" (1855) is thought to be the first fully-developed portrayal of such a creature. The story is headed by Melville with an anonymous epigraph, the third paragraph of which reads, "Seeking to conquer a larger liberty, man but extends the empire of necessity." In the story itself, a "great mechanician," Bannadonna, is commissioned to construct a huge bell-tower; after the tower is completed, he insists on working in secrecy on the belfry, eventually having a large object, concealed in wrapping, hauled up. Bannadonna alone remained in the belfry when the day came to inaugurate the ringing: the entire population remained below, but at the appointed hour, instead of the anticipated booming of the great bell, only a single muffled sound was heard, followed by silence.

Upon entering the belfry, the town magistrates found the dead Bannadonna and standing over him an enormous mechanical figure, cast by its creator to run upon a track at each appointed hour and strike the bell with its arms. Bannadonna, intent on

some finishing touches to the bell, had forgotten the hour and had been struck dead by the mechanical figure.

Yet, according to the story's narrator, this was to have been only the prototype for Bannadonna's ultimate creation, an "elephantine helot" to be produced in great numbers and incorporating all the characteristics of all the animals that mankind had heretofore yoked to its will: "All excellences of all God-made creatures, which served man, were to here receive advancement, and then to be combined in one." And the figure itself was to epitomize the aesthetics of the sublime: Bannadonna's design principle for it was "the more terrible to behold, the better."

Bannadonna had intended to give his "metallic agent" not only the power of locomotion but also "the appearance, at least, of intelligence and will." The terror inspired by the physical appearance of the automaton has its source in a deeper dread, originating in its violation of the border between life and death: inorganic matter, becoming animate by a process of purely mechanical or chemical operations, inevitably produces a reverse effect and draws the living into the realm of the dead. This is the third and final level of root metaphors about the machine.

In "The Paradise of Bachelors and the Tartarus of Maids" (also 1855), Melville casts the relation between humanity and machinery in these terms. The story's unusual structure is especially interesting, for Melville portrays the degeneracy or sterility of machine-based civilization not by contrasting it to a healthier, unmechanized condition but rather by juxtaposing it to another kind of sterility represented by traditional culture. The result, while wholly negative in tone, seems to make the point forcefully that there is no succor there.

The "Paradise of Bachelors" section recounts a long and very alcoholic dinner enjoyed by an old group of bachelors in an elegant private club in London; the story then shifts without transition to the "Tartarus of Maids" section, which describes a paper-mill factory in New England that employs a work-force made up only of young women. Both are based on visits by Melville, the first at Elm Court in Lincoln's Inn in 1849 and the second at Pittsfield, Massachusetts, in 1851.

The “Paradise of Bachelors” is a scene of sedate, well-tempered pleasure. The meal itself, although consisting of many courses, is curiously undistinguished fare; the dominant imagery is of the bachelors’ carefully modulated consumption style: not a one sneezes when the snuff is passed around. The meal itself is, as Dillingham remarks, “a metaphor for their orderly existence.” The impression of sterility and lifelessness is transmitted both by their dispassionate overindulgence in food and drink and by the state of lifelong bachelorhood to which all are committed.

The whole story’s structure – the abrupt succession of the two sections – employs the first as backdrop for the second. The intrinsically powerful imagery of sterility and death in the second section is heightened further by being presented against what had preceded it. The latter section is saturated with such imagery: the narrator-traveler’s close brush with death, the pallor in the female workers’ faces, the blankness of the paper, the factory (“like some great whited sepulchre”), the setting: “The mountains stood pinned in shrouds – a pass of Alpine corpses.” The traveler sees the apparatus inside the factory: “Something of awe now stole over me, as I gazed upon this inflexible iron animal. Always, more or less, machinery of this ponderous, elaborate sort strikes, in some moods, strange dread into the human heart, as some living, panting Behemoth might. But what made the thing I saw so specially terrible to me was the metallic necessity, the unbudging fatality which governed it.”

It is not just that the machine is the living entity; procreative allusions indicate that it has assumed the generative capacities of life as well. The machine is housed in a room that is “stifling with a strange, blood-like abdominal heat”: and the elapsed time between the introduction of the pulp and the emergence of the finished paper is “nine minutes to a second.” The female workers are all unmarried virgins whose very substance drains away. The traveler sees, imprinted on the finished paper, “glued to the pallid incipience of the pulp, the yet more pallid faces of all the pallid girls I had eyed that heavy day.”

The references to the "necessity" and "fatality" of the machine reinforces the epigraph to "The Bell-Tower": There is no escape from necessity through machine technology; on the contrary, that way leads to greater bondage.

One can assume that for Melville the world outside the machine's orbit was still vibrant and that no irreversible commitment to it had been made. By the end of the nineteenth century it seemed to many that such a commitment indeed had been extracted from a society seemingly enthralled by the system of machinery, especially in North America. The dominant opinion seemed to be that whatever unease the machine might evoke paled into insignificance beside the more immediate dangers against which man and machine warred side by side: the power of untamed nature, wilderness, and the surviving remnants of savage cultures. There is a marvelous representation of this attitude in the Currier and Ives lithograph *Across the Continent* (1868). A train is drawn up before a rough frontier settlement, on the other side of which two mounted native warriors stand; the train itself is the protective hedge for civilization against the as-yet-untamed wilderness.

Early twentieth-century imaginative fiction recognized this complete commitment (or capitulation) to the machine. The external form of representation that characterizes the first and second levels of root metaphor – the machine confronting mankind as master/servant or as automaton – gave way to imagery of full internalization. Portrayed in its most striking terms, the man/machine symbiosis emerged fully developed, with the inevitable result: degeneration of the physiological and psychological autonomy of the human agent. The machine appeared as metaphor for a human society organized as a single, machine-like organism.

E. M. Forster described his short story "The Machine Stops" as "a counterblast to one of the heavens by H. G. Wells." The human population resides underground, living singly in compartments where, at the pressing of buttons, mechanical devices supply water, food, air, beds, medicine, music, and communicating devices. Travel outside the compartments, although provided for, becomes rare, with a resultant deterioration in skin and musculature. Vashti, the central character, is described as a "swaddled lump of

flesh" with "a face as white as fungus." Originally the interlocking, supportive mechanism that sustains life in the compartments had been directly superintended by its designers; as the dependence became habitual, however, the human agents seemed to lose control over the functioning of the apparatus, which also had been supplied by its inventors with self-repairing mechanical aids. Soon they began to pray to it. That was the beginning of the end: "But humanity, in its desire for comfort, had overreached itself. It had exploited the riches of nature too far. Quietly and complacently, it was sinking into decadence, and progress had come to mean the progress of the Machine."

Eventually the mechanism collapses, taking with it the compartmentalized inhabitants. But they were already dead in all but name, the living dead. Kuno, Vashti's son, had tried to explain this to her before the end: "Cannot you see … that it is we who are dying, and that down here the only thing that really lives is the Machine? We created the Machine to do our will, but we cannot make it do our will now. It has robbed us of the sense of touch, it has blurred every human relation and narrowed down love to a carnal act, it has paralyzed our bodies and our wills, and now it compels us to worship it. The Machine develops – but not on our lines. The Machine proceeds – but not to our goal. We only exist as the blood corpuscles that course through its arteries, and if it could work without us it would let us die."

Hope for regeneration lies only in the rude bands of escapees or natives who exist completely outside the orbit of mechanical society. This theme recurs in *We* and *Brave New World*.

In *We*, the individuals – who carry such designations as D-503 and I-330 – are described as the "cells" of the "single mighty organism" that is the One State. All live in identical rooms and are nourished by a single, industrially produced substance. The Table of Hours regulates all movements, setting prescribed times for eating, work, exercise, and sleep, except for the two Personal Hours each day – which, it is expected, will soon become part of the "general formula" like the others. Zamyatin's imagery is dominated throughout by mathematical allusions. According to the sexual law, for example, each "number" (individual) is entitled to have sexual relations with any other:

"You declared that on your sexual days you wish to use number so-and-so, and you receive your book of coupons (pink). And that is all. Clearly, this leaves no possible reasons for envy; the denominator of the happiness fraction is reduced to zero, and the fraction is transformed into a magnificent infinity."

Society itself is a machine, an organism of differentiated and smoothly integrated component parts. A mechanism in the usual sense, the physical object, appears in *We* only as a symbol: first, as the Integral, a spaceship designed to bring the message of "mathematically infallible happiness," achieved by the One State, to other planets; and second, as the Benefactor's Machine, a device to cauterize the area of the brain that houses the faculty of imagination. The *One State Gazette* announces to the citizenry: "Until this day, your own creations – machines – were more perfect than you … The beauty of mechanism is its rhythm – as steady and precise as that of a pendulum. But you, nurtured from earliest infancy on the Taylor system – have you not become pendulum-precise? Except for one thing: Machines have no imagination … The latest discovery of State science is the location of the center of the imagination: a miserable little nodule in the brain of the pons Varolii. Triple x-ray-cautery of this nodule – and you are cured of imagination – forever. You are perfect. You are machinelike."

As the novel ends, D-503, chief mathematician for the Integral project, submits voluntarily to the operation: "It is the same as killing myself – but perhaps this is the only way to resurrection. For only what is killed can be resurrected." Once the operation is universally performed, and the imaginative faculty is genetically blocked in future generations, the mechanism itself will be needed no longer: society-as-machine will have removed all remaining impediments to its smooth functioning and will be able to reproduce itself identically for all time to come. But, at the city's edge, there is chaos, as the remnants of older humanity assault the surrounding Wall.

The matter-of-factness that Veblen identified as the behavioral orientation of the machine age has today become the expected routine of everyday life. We are accustomed to quantitative measure in every aspect of social life. The calculation of benefits and costs in numerical terms pervades our lives – in negotiations between

prospective marriage partners as well as between unions and corporations, in setting minimum levels of welfare payments as well as maximum "throw-weights" for nuclear missiles. Domestic life is unimaginable anymore without mechanical devices, and more and more people carry around inside their bodies some testimony to the wizardry of medical technology.

As well, an abundance of automatons in all sorts of horror films and science-fiction literature during the last fifty years has inured us to them; the ubiquitous video games should dissolve whatever remains of the machine's threatening visage. A few scattered souls may still quake at the prospect of self-programming computers becoming obstreperous, or of chess grandmasters being humiliated by an unanswerable gambit from a machine opponent, but for most the terror and dread, as well as the sublimity, that fired the nineteenth-century mind are gone. The relation between mind and machine is now grist for esoteric philosophical debate in the academic mills; the combat in this zone, however fierce it may become, is unlikely to revive that older mood.

The master of the new style is the Polish writer Stanislaw Lem, and the mode of representation is whimsy. *Mortal Engines* introduces us to "electroknights" and to "ultradragons" and to a computer that calls itself "Digital Grand Vizier" and insists on being addressed as "Your Ferromagneticity." *The Cyberiad* opens with a story about a machine that suffers with good grace the ridiculous commands of its inventor, although it cannot resist a touch of spite. The stories are infinitely comforting, because Lem's machines have all the pathetic emotions and foibles so readily recognizable as our own. And, after all, Jean Tinguely's self-destructive machine was designed to show precisely that the machine shares with us life's essential attribute, namely mortality, and is thus an affirmation of life rather than its negation.

Chapter 3: Modern Science and its Spacetime

> A Frenchman named Chamfort, who should have known better, once said that chance was a nickname for Providence. It is one of those convenient, question-begging aphorisms coined to discredit the unpleasant truth that chance plays an important, if not predominant, part in human affairs. Yet it was not entirely inexcusable. Inevitably, chance does occasionally operate with a sort of fumbling coherence readily mistakeable for the workings of a self-conscious Providence.
>
> Eric Ambler, *A Coffin for Dimitrios* (1937)

THIS IS A STORY BASED on the discoveries of modern science, particularly in twentieth-century physics, and the question to be posed at its ending is: How could anyone living in the modern age possibly find solace in such a tale when lying on one's deathbed?

In this tale, "reality" consists of three elements, although as we shall see, the largest portion of the reality of the universe as described by science is, at least for now, mysterious! Only a mere 4% or 5% of reality (a.k.a. the known universe), the portion that is known as mass-energy, making up what is currently detectable by us on earth and in the universe beyond, is well-described in scientific terminology. The rest is only inferred – that is, hypothesized indirectly – from our observations of the behavior of matter in the visible universe. Even the apparent reality of the matter all around us is misleading, for when modern physics writes "m" in its equations, the reference is to mass, not matter. For example, in the equation $W=mg$, the weight of an object on earth is said to be equivalent to its mass times gravity. In terms of any physical object, what we see with our naked eyes and think of in common-sense terms can be called matter, but the physical reality of that object is more accurately expressed as mass, which is invisible to us. All matter has mass, but so do many forms of energy.

Matter and energy are convertible: The "solid" matter we now know as being composed of atoms was once formed out of energy (around 400,000 years after the Big Bang), and in the great super-hot furnaces inside stars such as our sun, some matter is being turned back into radiant energy before our very eyes, although a basic postulate holds that the sum total of matter/energy always remains the same. In every dimension

of the visible universe there are well-grounded quantitative estimations of magnitude, including the origins of the universe itself and both the larger and the smaller aspects of its constituents: time, space, and the quanta ("packets") of matter/energy itself.

As to space: The diameter of the universe is at least 93 billion light-years across and the universe is still expanding. By the way, it makes no sense to ask what the universe is expanding "into," since the universe is by definition everything that exists in the space that is observable. Light travels at a speed of 300,000 kilometers per second, therefore one light-year measures 10 trillion kilometers in distance. Thus, the diameter of the universe is about one septillion [trillion trillion] kilometers, or 1 followed by 24 zeros. The 5% of visible mass-energy that exists in space is organized into 100 billion galaxies, like our own Milky Way, containing many trillions of stars, totaling somewhere between a sextillion [10^{21}] and a septillion in number.

As to time: The age of the universe is 13.82 billion years – actually, 13.799±0.021 billion (10^9) years – a length of time which we can at least crudely represent to ourselves: If one human generation had succeeded another during each period of 25 years since the beginning of time, then 600 million generations of humankind would have come and gone. But at the "other end" of time, namely its briefest dimensions, events apparently happen in units of duration so small as to be literally unimaginable. An optical atomic clock can measure time to one-quadrillionth of a second. Some events occur among subatomic particles – for example, the emission of a gluon from a quark – on a time scale of a *yoctosecond*, that is, one septillionth, or one trillionth-trillionth, of a second [10^{-24}s, written as 1 over 10 followed by 24 zeros]. The phase transition in the development of the universe that is known as "inflation," which was the onset of a process of exponential expansion, occurred from 10^{-36} seconds after the Big Bang to sometime between 10^{-33} and 10^{-32} seconds thereafter. Put into words, we are talking here about a timeframe that amounts to various fractions of one trillion-trillion-trillionth of a second. According to the theory of inflation, the Universe grew by a factor of 10^{60} in less than 10^{-30} seconds.

As to matter or mass: In 2013 the Planck Satellite, designed to measure the background cosmic radiation that is a legacy of the Big Bang, gave the following figures for the constituents of the universe: 5% is "ordinary" matter/energy, observable by us as stars, planets, galaxies, and so forth. A further 27% is "cold dark matter" – matter than cannot be observed by the usual "signal" of its electromagnetic radiation – which we presume on theoretical grounds must exist, although we don't know what it actually is made up of; its existence is inferred from its gravitational effects on matter in deep space. By far the largest share of the total, a whopping 68%, is referred to as dark energy, and again, with respect to what exactly we think this too is, we are pretty much clueless; its existence is inferred from the rate at which the universe is expanding.

And yet, when we turn from the external vastness of the universe and peer into the "insides" of all matter, we find, almost entirely, just empty space! The two building-blocks making up the nucleus of atoms, the particles known as the proton and the neutron, are 10^{-15} meters in size, that is, a fraction of a meter written as 1 over 10 followed by 15 zeros. The proton's mass is $1.67262158 \times 10^{-27}$ kg; the diameter of a proton is .85 fm (femtometer, which is one quadrillionth (10^{-15}) of a meter. The electron

is a kind of ethereal entity, far smaller than a proton, with a mass that is calculated as 9.1×10^{-31} kg [written as 9.1 over 10 followed by thirty-one zeros]. But whereas the electron is thought not to be made up of any subcomponents, both protons and neutrons are themselves composed of much smaller subatomic units known as quarks – in other words, their own interiors are mostly empty space.

Quite possibly there are still tinier units within: In what is known as string theory, the size of the oscillating strings that are hypothesized as being the ultimate foundation of all elementary particles are conceived in terms of a unit known as the "Planck length," which is 1.616199×10^{-35} meters; this is equivalent to 10^{-20}, or one sextillionth, of the diameter of a proton, and one of the scientific commentaries describes it, with perhaps unintended irony, as "an extremely small length."

Dimensions in time and space of such ludicrously small and large magnitudes cannot be represented concretely in the ordinary human imagination. The ultimate reality that is portrayed in contemporary astrophysics is one where the basic units of time and space are, paradoxically, both so immense and so vanishingly small as to defy the ability of most of us to think about them. At the tiniest levels, these invisible particles and forces produce a universe of truly vast dimensions, only a miniscule portion of which we can see with our naked eyes. Even the simple concepts of time and space – actually, for science, the unified concept of spacetime – have an aura of irreducible mystery surrounding them: What does it really mean to be told that the more distant are the lights our telescopes pick up, the further back in time we are taken, that we are seeing with their aid stars and galaxies as they were billions of years ago?

Thus, in point of fact both the old monotheisms and modern physics, featuring the equally invisible realms of souls and subatomic particles, respectively, utterly baffle ordinary understanding. At bottom, both are equally incomprehensible stories and therefore both must be taken entirely on faith. In what respect do they differ then? Essentially, religion allows us to believe that the universe we inhabit was made *for us*; science does not. That is all.

The "Big Bang" in which our universe came into being marks the beginning of time (and thus one cannot ask what happened "before" it). And with the reference to this event we are again in the realm of the literally unimaginable. The mention above of the detectable contents of space (the 5%) suggests by extrapolation that the sheer total mass of the universe is rather large. The magnitudes involved are simply staggering, as is illustrated by the gigantic columns of dust and gas out of which stars are continuously being formed: For example, the region of currently active star formation known as the Eagle Nebula, some 6,500 light-years distant from us (featured in the famous Hubble Telescope shot known as "The Pillars of Creation"), which is only one of countless similar regions of space, contains columns of dust and gas rising from it that are 100 trillion kilometers high.

And yet it is thought that, at time of the initial singularity, at the instant before the Big Bang, the *entire* mass of the universe was compressed into a single point of infinite density and temperature so small as to be incomprehensible on a human scale. It is sometimes said that the dimension of the totality of mass before the Big Bang was "roughly" a million billion times smaller than a single atom – and remember, the

diameter of an atom is just a few trillionths of a meter! The apparently "solid" stuff of an atom is only the tiny constituents of its nucleus, that is, a ball of protons and neutrons, and the nucleus is 10,000 times smaller than the entire atom (which includes its ephemeral electrons). What this implies is that the apparent solidity of matter conceals a vast emptiness within.

In view of the intellectual puzzles already presented, it is perhaps advisable to skip over some other current scientific conjectures about the natural home in which we humans dwell, such as the idea that we ourselves and our surroundings are mere holographic projections of another reality, or that there is a near-infinity [10^{500}: 10 followed by 500 zeros] of universes, not just the one we think we occupy.

The underlying idea about the development of our universe through time in discrete stages, such as inflation and the appearance of the first galaxies about one billion years after the Big Bang, is that our universe *evolved* into its current state purely as a result of the operation of its own "natural" forces and their laws of behavior. The known characteristics of the mass-energy transformations that succeeded one another over time during the past 14 billion years are, in terms of the reigning scientific theory, sufficient to explain many – but by no means all – of the obvious characteristics of the universe we inhabit today. The fact that current theory and observations in astrophysics cannot account for 95% of the universe means only that they this theory is incomplete, not that it is simply incorrect: Typically, later and more complete explanations incorporate earlier ones as special or limited cases, as in the case of Einstein's and Newton's conceptions of gravity.

When a fuller explanation has been achieved, as it will be, its basic building-blocks will be the same as the ones already known: spacetime and the quanta of matter/energy. Any more complete theory must make new predictions about the behavior of matter/energy that can be measured by instruments and verified by repeated observations. And one must always look to the "bottom line": Whereas the account so far is acknowledged to be incomplete, it also explains well an enormous body of accumulated evidence obtained and verified by rigorous methods. *There is, in short, no seriously competitive alternative approach to explaining the nature and origins of the universe*. This universe described by modern science is self-originating and self-sustaining, and if it is not eternal, that is because there is no necessity that it should be.

The same goes for us. What we call "life" evolved on planet earth as a special case of the same matter/energy dynamics that created and sustains the larger universe: For example, our bodies are composed of atoms that have all been recycled countless times and that were originally forged long ago in exploding supernovae and neutron stars. Some probabilities were involved, but there was no necessity in the unfolding of the original chemical syntheses occurring spontaneously on our young planet that led eventually to the appearance of the eukaryotic cell (some two billion years after the earth's formation), on which all complex life-forms are based. Nor was there any necessity in the progression of those life-forms, over the succeeding two billion years, as is shown by the periodic great extinctions during the last 500 million years of the earth's history; among other possibilities, another random collision of a massive asteroid with the earth could have ended the whole experiment, possibly for good.

Nothing illustrates better the brute fact of chance in evolution than what is known as the "Cretaceous–Paleogene extinction event," some 65-66 million years ago, caused by the massive asteroid which created the huge undersea Chicxulub crater off the coast of Mexico. Both the asteroid impact itself and the other events it triggered, including volcanic eruptions and climate change, resulted in the extinction of something like 75% of all the species then existing, including all of the non-avian dinosaurs. Until this time there were no mammals larger than rats, since the emergence of larger species had been inhibited by the top predators, the terrestrial dinosaurs. The succeeding period is known as the Cenozoic Era, or "the age of mammals," which over time became the dominant animal group on earth. Had this asteroid missed the earth, or collided with it much later, all of subsequent mammalian evolution would have been different, perhaps radically so, and modern humans might never have come into being.

In the later stages of mammalian evolution, neither the appearance of various *Homo* species out of our common ancestor with the chimpanzees, five to seven million years ago, nor the more recent success of *homo sapiens,* over the course of the last 300,000 years, in out-competing various descendants of *Australopithecus* and of other predecessors such as *homo erectus* and *homo heidelbergensis,* including our close cousins the Neanderthals, was inevitable. The appearance on earth of the marvelous intelligence that fashioned this scientific story was a chance affair. And someday it will disappear again.

The modern scientific basis of this story is inherently linked with technology, for were it not for continuous improvement in measurement instrumentation, in experimental and analytical methods, and in the composition of materials, scientific advance would have soon ground to a halt. In this approach there is a simple rule: Whatever is said to exist must have a magnitude (mass/energy) that can be measured and whatever is theorized must generate predictions about observations that can be made. The most famous example in the public mind was the experimental proof first obtained in 1919 for the bending of light-rays near massive bodies in space, a prediction derived from Einstein's equations of general relativity.

Another famous example is the Higgs Boson, long theorized in the standard model of subatomic physics as the particle that lends mass to matter; the theory also predicted (within a range of values) what its own mass must be. But until it was observed – finally in 2012/2013 – in the ghostly evidence collected by the powerful and extraordinarily complex machines known as particle colliders, which smash subatomic particles into each other at velocities close to the speed of light, its existence (and the viability of the entire standard model itself) was in question. The Higgs Boson example illustrates well the mediating role of technology with respect to advances in modern science: Theory and conjecture, in seeking to drive forward the process of new discoveries about nature, set continuous challenges for the development of novel instrumental and analytical technologies that are capable of making the observations and measurements needed to confirm the theories. Until the new technologies come onto the scene science cannot advance.

It is easy for the happy consumers of new technologies to be distracted by their gadgets and thus fail to notice that the reality described by science is not a very hospitable

place, all in all, especially for a creature inclined to worry, even just a little bit, about life after death. Science's universe is, in fact, mostly just cold dust and hot gas, spread across a space so vast – by comparison with the size of the human form – as to be literally unimaginable. The nearest star beyond our own Sun is Alpha Centauri, which is a bit more than 4 light-years (40 trillion kilometers) away. Whether there are other habitable planets out there, capable of sustaining life-forms similar to ours, is unknown, perhaps even unknowable, but it is likely, just on the basis of probabilities. However, given the distances involved, it is extremely unlikely that we will ever learn of the fact and less likely still that we will be in contact with their inhabitants.

Even if we did: So what? The universe we share with them will eventually suffer one of two fates: It will either grow frightfully cold, dark and lonely, through accelerating expansion, or the expansion will stall and reverse itself, whereupon a furious heat-death will consume everything in it. Life as we know it is the rarest thing in the universe and when all the inhabitants of planet Earth vanish, as they must one day, it will be rarer still. By one billion years from now our sun's own evolution, dictated by the physics of stars, will have caused it to grow hotter, hot enough to boil away all liquid water on earth and bake the ingredients in the earth's crust into a solid metallic sheet.

Even if some human descendants are still around to witness the event – an unlikely prospect for such an aggressive and insecure ape as we are – the sun's searing heat will mark the end of this relatively brief experiment with life in one small corner on the fringes of the universe: not with a bang, but with a whimper. All without ever having had any evident purpose or meaning; without any plausible reason to think that we humans are "special" in any way (except in our own estimation); with no particularly remarkable result, just a recycling of the atoms formerly constituting human bodies and all the cultural artefacts they crafted into alternative molecular configurations; gone without a trace, with nothing at all left to mark the past, present, and expected future of human civilization, except the strange capsules we once propelled into the void of empty space, proclaiming our wish for contact with someone, anyone; without a hope or a prayer, just a deep satisfaction that for a short while we had been blessed by nature with a brain of such remarkable power and ingenuity that the extraordinary complexity of the universe which gave birth to us stood at least partially revealed before it.

This perspective is far too bleak for most people, even if they don't think they could do without the technological blessings wrought by science. But in a real and ironic sense it also shares with its religious counterpart a deeply mysterious character, despite its hyper-rational mode of representation. On the simplest, intuitive level, the idea that the matter of the earth we experience as reassuringly solid is, in point of fact, mostly a vast empty desert is almost impossible to grasp in any practical sense.

Consider also the neutrino – a vanishingly tiny particle with exceedingly small but nonzero mass – as it whizzes unimpeded, at velocities close to the speed of light, trillions of them at a time, straight through planet earth: right through our bodies, through the lithosphere (including the crust we stand on), the asthenosphere, the mantle, the outer core (liquid iron), and the inner core (solid iron), and out the other side again, *without touching anything (except extremely rarely)*. Its path through the earth was once described by a commentator on a BBC scientific program as being "like a

bullet passing through a bank of fog." It is impossible for most of us to imagine just what this particle actually *is*.

Every one of the magnitudes that modern physics refers to – the speed of light, the dimensions of the universe, the size of subatomic particles, the yoctosecond, the idea that in the instant before the Big Bang everything in the universe could be condensed into a single point of infinite density far too small to be imagined by our ordinary brains – defies imagination. The scientific conception of reality simply blows away common sense and the capacity of understanding possessed by the vast majority of human beings who have ever lived or will live in this universe.

And it gets worse, for when physicists get down to work, they do not even use ordinary language, or any language familiar to most persons, when expressing their thoughts. Instead, they rely on a symbolic mathematical/geometrical notation that is itself, to most of us, no more comprehensible than would be the ancient Semitic language that Jesus of Nazareth spoke in everyday life (Aramaic) or the Greek used by the authors of the Gospels – which Jesus probably wouldn't have understood – when they were making up the New Testament many decades after his death. (The conceit that modern-day English-speaking evangelicals think they know the literal "meaning" of a Biblical text is hilarious, but that's neither here nor there at the moment.)

Here is a famous example of scientific notation, from the Wikipedia entry on Einstein's field equations of general relativity, which may be written in the form:

$$R_{\mu\nu} - \frac{1}{2}g_{\mu\nu}\,R + g_{\mu\nu}\Lambda = \frac{8\pi G}{c^4}T_{\mu\nu}$$

where $R_{\mu\nu}$ is the Ricci curvature tensor, R the scalar curvature, $g_{\mu\nu}$ the metric tensor, Λ is the cosmological constant, G is Newton's gravitational constant, c the speed of light, and $T_{\mu\nu}$ the stress-energy tensor.

In these equations are described the fundamental nature of the four-dimensional spacetime we apparently inhabit. It is likely that only a tiny fraction of all the humans who will ever walk this earth would ever know what is meant by them (although a few will at least recognize the numbers 8 and π). But at least we have John Wheeler's elegant and concise translation to help us: "Spacetime tells matter how to move; matter tells spacetime how to curve."

Is the type of belief we hold about the "truth" of modern science any different from the belief in a spiritual reality (Jewish, Christian, or Muslim) or, for that matter, a belief in elves and goblins?

Of course, no one is precluded from believing both stories and many profess to do so, although certain inescapable curiosities present themselves immediately in such an

exercise. For example, what was the creator-deity doing for all those years before the appearance of these precious humans? To be precise: Assuming that the origins of Judaism are to be found in the first millennium BCE, how was the deity occupying its time for the approximately 13,800,000,000 years prior to that? In fact, the timeframe for the origins of all three monotheisms fits well within the margin of error for the dating of the universe's formation (13.799±0.021 billion years ago). And why create so much space, if the only drama of interest in the whole universe is the deity's constant hectoring of a few misbehaving mammals which were set down on a pathetically small hunk of rock on the outer edge of a quite ordinary galaxy?

True, an ordinary mortal is not allowed to interrogate the deity, especially about its plans and purposes – only Satan gets away with that – but, still, it's all a bit puzzling. And why wait for the Jews to figure out the divine plan, which leaves all those human souls created in earlier millennia sitting in limbo? (Not even the Catholic Church can any longer figure out how limbo is supposed to work.) Perhaps the only good answer is that for an eternally-existent being 14 billion years goes by in the blink of an eye. Still, it is a bit disappointing that the deity didn't wait for just a little bit longer to reveal Itself to all of us, until the Internet was available, so that we could all subscribe to the Divine Blog and Tweets – except that, like the Government of China, the established churches would be loath to allow the deity to have direct, uncensored communication with individual believers.

Belief in both is relatively easy for the scientist-religionist: One can simply affirm that the deity designed and enacted the natural laws by means of which the universe as a whole, and we in particular, evolved – although why it took so long is none of our business. True, the deity's existence violates the dictum of Occam's Razor, which advises us "not to multiply entities unnecessarily," because it just replaces one insoluble mystery (where the point of infinite density before the Big Bang came from) with another (where the deity came from). This minor conundrum is easy to tolerate since it allows one to have a healthy serving of both scientific truth and personal salvation at relatively little incremental cost in terms of time and resources.

It is well known that the Catholic Church tried mightily to maintain control over the interpretation of scientific results, until by the nineteenth century social changes forced it to relinquish its hold. Less appreciated is the common enterprise that has bound them together. A famous thesis, associated with the German sociologist Max Weber, refers to the "disenchantment" of the world resulting from the rise of capitalism, science, and the modern state. The human world was once "enchanted" by and with a multiplicity of spirits – ghosts, goblins, elves, fairies, and all the rest. They were driven out first and foremost by the increasing grip of monotheism and its rationalistic theology, which put the rich and vibrant spirit-world of earlier days on a strict diet: Islam, for example, has in addition to Allah only jinns, angels, Iblis (Satan) and lesser devils as insubstantial entities; Christianity, just the three-person Godhead, angels, saints, Satan and his helpers. For both all other spiritual entities are strictly proscribed by the all-powerful Allah/God. Then modern science came along to finish the job, as immortalized in the reply of the great French mathematician Laplace to Napoleon, who queried the absence of a single mention of the creator-god in his five-volume

mechanistic account of the universe's operation: "Sire, I have no need of that hypothesis."

Although modern science collaborated with its early antagonist, Christian monotheism, in ridding the world of most of its traditional spirits, in so far as the capabilities formerly attributed to them were among the presumed causes of natural events, it did not put an end to nature's mysterious aspects. Indeed, the odd thing is that, as the sciences progress, driven onwards by their hyper-rational methods of investigation, the physical world around us seems to get progressively *more* mysterious, not less. Some of these mysteries have to do, as noted above, with the bizarrely tiny dimensions of the ultimate constituents of matter itself. Others appear when the "solidity" of matter vanishes as we smash it to smithereens in the huge particle colliders. The bizarre behavior of sub-atomic particles as described and verified in quantum mechanics (entanglement, superposition, wave-particle duality, etc.) is another source.

Most of the mysteries that defy representation in our imaginations come out of sub-atomic physics, to be sure. There are many others, which might be called marvels rather than mysteries, which challenge the common understanding of the world. In chemistry one thinks, for example, of the three-dimensional self-assembly of the large, complex molecules known as proteins that generate all of our bodily functions. In molecular biology and genetics, the infinitely complex processes involved in information transfer within the genome; in atmospheric science, the modelling of climate system dynamics with many key variables over long time-frames; in medicine, the interaction of myriad risk factors for diseases. One should not even mention the ghostly entities that are played with in higher mathematics.

Most people understand that religion and modern science offer competing explanations of important phenomena by using radically different methods. Most do not expect or require religious dogmas to rest on the same evidentiary standards (in particular, replication of experimental findings) which scientists employ – or so they have heard, at least. The religious mysteries have been around a lot longer, of course, and for the most part are not hard to grasp. We can all imagine a bearded old white man whipping up the solar system and giving life to Adam (just look at the ceiling of the Sistine Chapel); we have seen the pictures of the beam of light entering Mary's ear, explaining the virginal conception; we can all appreciate the attractions of paradise available to the Islamic warriors.

All the hard stuff has vanished into the dim past. How many among the faithful today know that it took centuries for the first Christians to suppress widespread dissent around the dogma of their weird three-part deity, the Triune God – which to the other two Abrahamic faiths, Judaism and Islam, proper monotheisms both, is just a thinly-disguised recrudescence of polytheism?

On the other hand, most of us cannot even imagine how to represent in our minds the mysterious world of matter at the subatomic level: Even reputable quantum physicists have described it with words such as "surreal" and "imaginary." *We cannot visualize this world* (as Einstein complained). We *can* visualize both particles and waves, which is where quantum physics started, with the concept of the photon, the basic unit of light energy, which behaves sometimes like the one (particle) and sometimes like the other

(wave). Alas, that was not the end of the matter, for it was later discovered that all subatomic particles behave in similar fashion. Then scientists found that there are not two discrete and different states, but rather a continuum, that is, a continuous spectrum between the two separate states of matter at the smallest scales. Nothing "solid" remains for us to anchor our imagination onto. Our home, the universe, a physicist explains, "is not made up of particles and waves and beams of light with a definite existence … [but] is aware of all the possibilities at once and trying them out all the time." Even though that other experiment, where one tries to improve the clarity of one's thinking by battering one's head against a brick wall, would appear to be a conclusive demonstration of the contrary.

What are we to do with this fecund science? In a very real sense, we have no choice but to take the whole corpus of generally-accepted science on trust, in part because of the public character of the scientific enterprise. To put the matter bluntly, that enterprise is always evolving, and every new answer is always provisional – but it is simply unthinkable to really imagine that the whole enterprise could be some kind of elaborate hoax. At the simplest level we can all see the evidence in front of our eyes, every day, namely, that our gadgets and medical therapies work (within acceptable parameters of performance). We even have a firm intuition that all of it goes back to work in basic sciences that most of us cannot grasp: No CD-player without laser beams, no lasers without photonics, quantum theory, Einstein's pathbreaking theoretical papers of 1905 and 1917, and subsequent technological development.

Once in a great while the idea that all or some of modern science might be a hoax rears its head, the most notorious example being the recruitment of a few Nobel-prize physicists in the early years of the Nazi party in Germany in support of the calumny that Einstein's relativity theory was an expression of "Jewish physics." The scientific community itself roots out such nonsense efficiently and definitively.

There is more that needs saying about what old monotheism and new science have in common. Both share this curious feature, namely, that the world we apprehend with our senses and ordinary understanding is not the "real" world, but rather an illusion. Although the real is the "first cause" of what we actually experience, in the sense that it makes the reality we apprehend possible, *the real itself is invisible*. For monotheism, what is real are called "spirits," that is, types of disembodied entities, such as souls and devils, some of which can assume visible physical form on occasion, notably in the life of Yeshua (Jesus) on earth. For modern science, the "forces" by means of which everything happens in the 5% of reality that is the visible universe –electromagnetism, gravity, the strong and weak nuclear forces – are entirely hidden from view and have no physical form. (Hidden away yet more cleverly still are dark energy and dark matter, making up the other 95%.)

And the subatomic particles that are the ultimate building-blocks of the matter we can see are themselves visible – for tiny fractions of a second too small to imagine – only as some ghostly trajectories recorded as they are decaying after being smashed into smithereens in the particle colliders. If vibrating "strings" turn out to be the very first level of material reality, one can be virtually certain that we will never actually see one. In this respect monotheism and science alike are content to infer the existence and

qualities of their invisible realms through deductive reasoning from both theory and analogy with observable things.

Our contemporary secular society, so beholden to the continued progress of the sciences to feed the insatiable maw that is our need for technological innovation, simply has no choice but to take the truth of scientific findings on trust. (A curious exception is climate science.) The need for faith is thus no less intense now than it was in the Middle Ages; only the object of faith has changed. The vast majority of the believers who are dependent on the institutions which guide faiths and their key personnel – first priests, now scientists – find themselves in a very weak position. The mysteries of faith are a closed book they cannot comprehend. The responsibilities of those who are supposed to guide them along the one true path are correspondingly enormous, for they know, or ought to know, that their followers have little choice but to take their words on trust. Yet how often in the history of monotheism do we find the simple substitution of lethal force for gentle persuasion: hand in hand, the Holy Book and the sword? The human agents who run the institutions of monotheism could never quite put their trust in the voluntariness of belief.

To ordinary human understanding, science's accounts of the unfathomable mysteries about how the universe functions are, if anything, far more obscure than the contents of the densest theological treatises. Those who guide this enterprise have therefore an even higher level of responsibility toward the uninitiated than priests do. But it turns out to have been near-impossible to specify the nature of the type of responsibility that scientists bear toward the rest of us, in large part because most members of the scientific community have never had the slightest interest in elucidating this point for us. Rather, they have been obsessed with an inward-looking form of responsibility for upholding the integrity of proper method: For example, scientific fraud is understood in these terms, as a deliberate betrayal of the accepted canons for hypothesis-testing, evidence-gathering, and good laboratory practices. The rest of us are not invited to meddle in these discussions.

So, our trust in science and scientists is well-placed. But the potential for the betrayal of trust is always there, and becomes ever more problematic as science advances. This potential has a number of aspects, the most important of which is the idea that the sole determinant of what is or is not legitimate research is that it should be conducted in accordance with the latest accepted protocols on methods. These protocols include ethics, of course, but limited almost entirely to the use of human subjects. With very few exceptions, scientists do not want to be asked to make any other judgements about the possible desirability of either foreseeable or unforeseeable applications of research results in the larger society, especially out into the distant future.

This strategy worked well for a very long time, although it almost broke down in the 1930s, with the juxtaposition of atomic science and totalitarian politics in Nazi Germany. It is unlikely to work very well in the future.

Chapter 4: Seven Figures and the Agony of Modernity

IMAGINE A HORIZONTAL LINE drawn along the 50° N latitude across Europe, extending from the westernmost part of Germany – where Germany, France, and Belgium meet – to the easternmost part of the Czech Republic. This line, which passes through the cities of Frankfurt/Main and Prague, should extend more precisely in terms of longitude from about 7° E to 18° E. Now place a vertical line extending south at the western end of the first line, to 48° N (where Munich and Vienna are found), thus encompassing two degrees of latitude, and complete the elongated rectangle, which in linear distance will run about 900km from West to East and about 300km from North to South. Highlighted on a map of Europe, this would appear on the vertical axis as a narrow band of territory stretching horizontally from the western border of present-day Germany to the eastern border of what is now Czechia (Czech Republic).

In the earlier part of the nineteenth century, when the story related here begins, the modern nations of Germany and the Czech Republic did not yet exist: What later became Germany was forged by Bismarck in 1871 out of some dozens of principalities and states in the German Confederation, and the lands now known as the Czech and Slovak Republics were a part of the Austro–Hungarian empire. Our rectangle encompasses some famous old cities, of course, but at that time it was largely composed of smaller towns and villages, populated by merchants and artisans and surrounded by small farms and larger landed estates. Within these little towns lived a cowed and oppressed people, of ancient lineage, which was deprived of full membership in civil society and whose children traditionally had been denied access to higher education and many of the ordinary occupations, especially professions such as medicine and law, but even farming – although these deprivations were beginning to change throughout the nineteenth century and would be substantially cancelled by the end of that century.

Figure 6 Map of Europe in 1848

These people were of course the Jews. They sought to live unobtrusively amongst their compatriots who professed various forms of Christian faith, while holding ancient memories of slaughters and torture during the Crusades and the Inquisition, and more recent acquaintance, further East in Poland and Russia, where their own populations were substantially larger than they were in the West, of periodic violent pogroms and confinement to the *shtetl*. Embedded deeply in both of the competing Abrahamic religion, Islam and Christianity, antisemitism had become a persistent, pervasive, poisonous vapor, hugging the ground like a heavier-than-air gas, until it seemed as if it has saturated nature itself so thoroughly that even the horror of the Holocaust has been unable to wash it away.

In Western and much of Central Europe the Jews were but a tiny minority of the whole population, and yet, out of this little cadre scattered throughout a relatively small

geographical area (our irregular rectangle), there emerged in the nineteenth century a set of seven individuals, drawn from the Jewish communities therein, all of whom would make ground-breaking intellectual innovations, revolutionizing art and thought and helping to inspire the creation of what we now know as modernity. These seven, in order of birth, were Karl Marx, Sigmund Freud, Edmund Husserl, Gustav Mahler, Albert Einstein, Emmy Noether, and Franz Kafka.

For my purposes, it is unnecessary to speculate overmuch on how or why such a flowering of individual genius could emerge from these small communities which were still living in the shadows of the larger society. One important factor, to be sure, was that during the nineteenth century in Europe those communities were being freed from centuries of civil restrictions on their activities and educational opportunities, leading them to full civic participation while by no means relieving them from the persistent and pervasive anti-Semitism among their countrymen. Another factor is the long tradition of love and respect for learning among them, embodied in their rabbis. There are surely other factors as well. What, in brief summary, was their life, work, and fate?

Karl Marx (1818–1883) was born in Trier (lat. 49º N, long. 6.6º E), at the westernmost edge of our rectangle, in what is now the southwestern German state of Rhineland–Palatinate. Marx's maternal grandfather was a Dutch rabbi, while on the paternal side his forefathers, down to his grandfather, had been the rabbis in Trier since the early eighteenth century. His upwardly-mobile father, however, who became a lawyer and owned some Moselle vineyards, converted from Judaism to Lutheranism before his son's birth in order to be freed from the many civil restrictions still imposed on Jews. Marx, who is, I believe, without peer in the field of modern social thought, cannot carry any personal blame for the misuses of his theories in the service of oppression in the Soviet Union, China and elsewhere. Among other things he had a great impact on progressive social democracy, especially in Germany, and early twentieth century European history might have taken a better course had his followers been able to resist the destructive nationalism that led to World Wars I and II.

Sigmund Freud (1856–1939) was born in what was then the small village of Freiberg – now called Příbor – which lies in northeastern Moravia, and is at the easternmost edge of our rectangle at lat. 49º N, long. 18º E. His ancestors were Hasidic Jews from Galicia (Ukraine) and his father, a wool merchant, was known as a scholar of the Torah. Sigmund, who was born in a rented room in a locksmith's house, was three years old when his family moved to Vienna. Although he is best known for psychoanalysis and the theory of the unconscious, in my view the great depth of his thought is best found in the area of socio-cultural theory, especially in the short work, *Civilization and its Discontents* (1930). By 1938 he had a worldwide reputation and was celebrated by many notable intellectual figures, which failed to impress the Nazi authorities, who held him for ransom in Vienna, where he was already suffering from the buccal carcinoma – the result of his lifelong cigar smoking – that would kill him little more than a year after his settling in exile in London. His four elderly sisters, who had been left behind in Vienna, were murdered in the Nazi camps.

Edmund Husserl (1859–1938) was born in the small village of Prostějov (Prossnitz in German) in eastern Moravia (lat. 49º N, long. 17º E), a short distance from Freud's birthplace. His father was a textile merchant. After early study in mathematics, he switched to philosophy and became known for founding the school of thought known as phenomenology, and his later illustrious academic career took place entirely in Germany. His work remains very influential down to the present day, with many later thinkers paying tribute to his insights, and it is fair to say that he was the most important philosopher of the twentieth century. His last and perhaps most important work is *The Crisis of the European Sciences* (1936); the original German text of this book had to be published in Belgrade, since by that time Hitler's regime had proscribed all publications by Jews (his earlier conversion to Protestantism made no difference, of course).

Gustav Mahler (1860–1911) was born in the village of Kaliště or Kalischt (lat. 49º N, long. 15º E), in eastern Bohemia, now in the western half of the Czech Republic. His grandmother had been a street peddler, and his father became a coachman and later an

innkeeper. When he was four years old his family moved to Iglau, a short distance south, where his father operated a tavern and distillery. At age 15, following a short and unhappy try at schooling in Prague, Gustav was accepted at the Vienna Conservatory of Music after auditioning before the famous pianist Julius Epstein. Mahler had converted to Roman Catholicism (the state religion of the Austro–Hungarian Empire) in 1897 in view of the prohibition against appointing a Jew as director of the Vienna Court Opera; shortly thereafter he was also appointed as the conductor of the Vienna Philharmonic Orchestra. As a composer, he was undoubtedly the greatest figure in the Western tradition of symphonic and vocal music in the twentieth century.

Albert Einstein (1879–1955) was born in Ulm (lat. 48º N, long. 10º E), in what is now the southern German state of Baden–Württemberg, at the southern edge of our rectangle. At the age of one year his family moved to Munich, where his father and uncle were industrialists who manufactured electrical equipment. There is a telling incident from his childhood in Munich: One of his teachers brought a long nail into the classroom, telling the students that with such a nail Christ was affixed to the cross, after having been betrayed by the Jews, whereupon all the other students instantly turned in their seats to stare at the young Albert. The five path-breaking papers he published in his *annus mirabilis* of 1905 – the result of working entirely alone as a clerk in the Swiss patent office – are a landmark in the history of physics. His special theory of relativity from that year was so revolutionary that the Nobel Prize committee, belatedly awarding him the physics prize in 1921 (but not for this theory), forbade him to mention relativity in his public speech at the awards ceremony – a prohibition which Einstein mischievously ignored. His general theory of relativity from 1916 remains today the unsurpassed account of the workings of the universe on a large scale.

Amelia Emmy Noether (1882–1935) was born in Erlangen (lat. 49º N, long. 11º E), a small city situated northwest of Nuremberg in Bavaria. Earlier generations of her family on both sides were merchants, but her father, who was largely self-taught, made a reputation in mathematics. She was called by Einstein and others the most important woman in the entire history of mathematics, and “Noether’s theorem” has been

described as the single most important mathematical contribution to modern physics. Despite her acknowledged brilliance, as well as the support of influential professors such as Hermann Weyl and David Hilbert, no university in Germany would award her a professorship on account of her gender.

Noether can appropriately stand as the representative for an entire generation of Jewish women who broke down the barriers to higher education in the sciences in those years, such as Lise Meitner (1878–1968), the first woman to become a full professor of physics in Germany, who discovered atomic fission and who was forced into exile in Sweden after 1938 and was unjustly deprived of a share in the Nobel Prize won by her long-term collaborator Otto Hahn in 1944. And Clara Immerwahr (1870–1915), born on a farm near Breslau in what was then eastern Prussia (now Wroclaw in Poland), who was the first woman to be awarded a doctorate in chemistry in Germany, and who, having married Fritz Haber, killed herself with his service revolver after failing to persuade him to stop his work on weaponizing chlorine gas during World War I.[1]

Franz Kafka (1883–1924) was born and lived his whole life in Prague (lat. 50º N, long. 15º E), in the state of Bohemia, Czech Republic. His family was Ashkenazi Jews, and his father was a successful salesman and retailer. Franz was diagnosed with tuberculosis in 1917, at the age of thirty-four, and spent most of the rest of his life in

[1] The case of Fritz Haber (1868–1934) illustrates many of the themes under discussion here. Born a Jew in Breslau (now Wroclaw), Haber converted to Lutheranism and became a strong German nationalist. He won the Nobel Prize in Chemistry in 1918 for developing the synthesis of ammonia from nitrogen and hydrogen; the Haber-Bosch Process created industrial quantities of synthetic fertilizer, and it is estimated that half of the human population is alive today due to its impact on agriculture and food production. (It also became a key component in high-explosive shells and bombs.) During World War I his nationalistic fervor prompted him to seek an officer's commission, then to weaponize chlorine gas and personally supervise its release against enemy soldiers on both the Western and Eastern fronts. In the early 1920s scientists working in Haber's lab at the Kaiser Wilhelm Institute developed a cyanide-based pesticide commercialized as Zyklon A; its successor, Zyklon B, was used to kill millions in the Nazi extermination camps. At the University of Berlin Haber and Einstein became close personal friends despite their sharply differing political views, but Haber dutifully voted to terminate Einstein's membership in the Prussian Academy after Einstein fled Germany in 1932. He himself lost his university professorship in 1933, despite Max Planck's attempt to intervene on his behalf directly with Hitler by recalling Haber's service to the German state (Hitler told Planck, "A Jew is a Jew"), and he died in exile in Switzerland the following year.

sanatoria, dying in Austria where he had gone for treatment. None of his major works had been published at the time of his death, and the unpublished manuscripts were in the possession of his friend Max Brod; Kafka had instructed Brod to destroy everything upon his death, but Brod thankfully ignored this direction. Kafka's early demise at least saved him from the fate of his sisters, three of whom were murdered in the extermination camps of Nazi Germany.

In summary, four of the seven (Freud, Husserl, Mahler, and Kafka) were born in either Bohemia or Moravia, and the other three (Marx, Einstein, and Noether) in southern and southwestern Germany. Their birthdates span a period of sixty-five years; leaving aside Marx, the span for the remaining six is a mere twenty-seven years. Five of the seven (all except Noether and Kafka) came from small towns or villages, and all were raised in German-speaking communities. *And a mere sixty years after the youngest of them was born – Kafka, in 1883 – all of the communities from whence they sprang, which had existed in many cases for more than five hundred years, were extirpated, root and branch, and their entire populations wiped out, in the maelstrom we know as the Holocaust.* The Nazis were known for nothing if not their methodical thoroughness.

As adults, most of the seven (all except Kafka and Husserl) became exiles from the lands of their birth. Marx, having moved around Germany, France, and Belgium for a number of years in the 1840s, harassed by various political authorities, finally fled to London in 1849, where he spent the rest of his life. Freud was still living in Vienna in 1938 when the *Anschluss* – the annexation of Austria by Nazi Germany – occurred, and was allowed to leave for London only after France's Princess Marie Bonaparte paid the exorbitant "flight tax" levied on him by the Nazi authorities. Persistent and vociferous anti-Semitism finally drove Mahler out of Vienna in 1907, when he moved to New York to conduct the Metropolitan Opera and the New York Philharmonic Orchestra; he fell seriously ill in New York in late 1910 and died almost immediately after arriving back in Vienna in early 1911.

Einstein had faced death threats and public denunciations in Germany throughout the 1920s. He was in Berlin in 1922 when his good friend Walther Rathenau, then

Germany's Foreign Minister, was assassinated, and he left the land of his birth for the last time in late 1932 to settle in the United States, just ahead of Hitler's accession to power, vowing never to return.[2] Emmy Noether, having been dismissed from her low-level teaching position at the University of Göttingen in 1933, also settled in the United States, where she tragically died from complications following surgery only two years after finding a safe haven. Husserl died in Germany in 1938, after having had his university privileges as a retired professor revoked in 1933; had he lived a few years longer, he would have been persecuted and perhaps killed, since old age was no barrier to torment for Jews in Nazi Germany – some too aged and infirm to leave their beds were carried from their homes, bedding and all, to the trucks waiting to take them to the camps. Had Kafka (who was only forty-one when he died in 1924) lived on for another two decades in his native Prague, would have been either forced into exile or murdered like his sisters.

Remarkably, five of the seven came from communities lying along latitude 49º N, and three of them – Freud, Husserl, and Mahler – were born only four years apart in villages stretching from eastern Bohemia to eastern Moravia and lying virtually in a straight line about 200km from each other. Some other interesting coincidences unite the seven. Marx and Freud both died in exile in London, albeit many years apart. Freud and Mahler grew up in Vienna, which by the late nineteenth century had become a hotbed of anti-Semitism under its notorious long-serving mayor, Karl Lüger. (In general, throughout the late-nineteenth and early twentieth centuries popular anti-Semitism was far worse in Austria than it was in Germany itself, and due to his high public profile later in life Mahler in particular suffered from it.) During his existential crisis over his marriage to Alma, Mahler travelled to consult Freud at the latter's vacation spot in

[2] A wonderful treasure from this period is the *Born–Einstein Letters*. Einstein's close friend Max Born (1882– 1970) was, like Haber and Immerwahr, a Jew born in Breslau (Wroclaw) and became one of the great geniuses of atomic physics during the 1920s, known among other achievements as one of the founders of quantum mechanics. Like so many others he left Germany after 1933 and became a professor at Edinburgh before returning to Germany in his retirement. He and Einstein never saw each other again after both emigrated, but thankfully their extraordinary wartime correspondence survives.

Leiden in southern Holland in the summer of 1910. Einstein spent the academic year of 1911/12 at the University of Prague, where he became acquainted with Max Brod, Kafka's faithful friend; we cannot be sure that Kafka and Einstein ever met, although Kafka did occasionally attend the meetings of Jewish intellectuals frequented by Brod and Einstein. Einstein knew and greatly admired Noether and her work, and both ended up in exile in the United States.

My story about this set of seven figures is used here in a symbolic rather than an evidentiary sense. Of course, there were other as yet unmentioned late-nineteenth-century pathways to modernity not represented here – in painting and sculpture, for example, as well as in theatre and dance; in many other disciplines in the human sciences; in the concepts of evolution and electromagnetism; in chemistry, geometry, architecture, and more. And there were many other major figures of great genius, non-Jews and Jews alike, who forged its key tenets in art and thought.

The century after 1850 in the Western world (Europe and North America) saw a stunning matrix of changes in economy (large-scale industrialization), politics, society, culture, and the sciences. It turned out to be a very mixed blessing indeed. For the matrix of changes had a double aspect: On the one hand, it represented a powerful amplification of the eighteenth-century French Enlightenment, with strong progressive currents in politics, society, culture, and the sciences (natural and human). On the other hand, there was a terrifying reaction against all these innovations, against democracy, against the new culture (called "degenerate"), against tolerance and individual freedom, against even modern science itself. That reaction was at first weak and faintly ridiculous, marked by the hysterical rhetoric and pathetic strutting of demagogues. But when it was finally consummated in Hitler's new order, there was revealed a "solution" to the hatred of modernity so harrowing, so brutal, so all-encompassing, and so pitiless that it has proved hard for succeeding generations to fully come to terms with its utter depravity.

To borrow a term from the German military theory of *Blitzkrieg*, what one looks for in a swelling movement of ideas that succeed in smashing the defenses erected

around older and sclerotic intellectual systems is the *Schwerpunkt:* the battering-ram, the concentrated mass of forces that achieves the decisive breakthroughs. The story related here suggests that the *Schwerpunkt* in all this stunning intellectual development occurred among the German-speaking peoples in Western and Central Europe. And it was the tragic fate of the Jewish communities there to embody, more clearly than any other identifiable ethnic grouping, that *Schwerpunkt* – to help mightily to lead the progressive forces, and then to suffer disproportionately the full, awful consequences of the backlash against them. The clash is best symbolized in Germany during the decade of the 1920s. The one side featured the growing power of Hitler's storm-troopers and his hateful propaganda. Meanwhile, that same nation was at the center of the most fateful new field in the natural sciences, atomic physics (where about one-third of the most important findings, and the Nobel Prizes that celebrated them, were attributable to Jewish scientists); and it was also the location of the city of Berlin, where the avant-garde in arts and culture had their wares on display.

Towards the end of the nineteenth century many physicists had believed that their discipline was fully completed, and that there was nothing of importance to be newly discovered. This was a typical expression of the pervasive cultural and intellectual smugness in the dominant culture of that time – and the political smugness as well, illustrated by the view that major wars were a thing of the past for the European nations. Yet just as the new century began to unfold, the seven figures profiled here, and their many contemporaries, were tearing down the intellectual and cultural edifices they had inherited. After another decade had passed, World War I shattered the prevailing social institutions and political structures of Europe.

Europe was at the center of the historical development of modernity, and the German- speaking world formed its core. The coming of modernity involved the dissolution of traditional and seemingly stable forms of thought and artistic expression that had been dominant for centuries prior, across a truly breathtaking range of forms:

- In cosmology, grasping the true and astonishing spatial and temporal dimensions of the universe;

- In physics, Einstein's relativity and the concept of space-time, replacing the intuitively more familiar Newtonian scheme;
- In subatomic physics and quantum theory, an entire world of unusual natural processes fundamentally at odds with the nature we apprehend with our senses;
- In geology, the vast time-expansion of the earth's biography and its profound changes in climate and physical topography;
- In psychology, Freud's theory of the subconscious and of a set of chaotic, hidden mental processes underlying personality, culture, and social dynamics;
- In popular music, jazz, and in classical music, the development of atonality;
- In painting and sculpture, the dissolution of familiar forms of the human body and the natural world in impressionism and expressionism;
- In architecture, the *Bauhaus;*
- In Kafka's prose, the undermining of the individual protagonist's autonomy within the obscure workings of bureaucracy;
- In evolution, the shattering of the myth of one-time human creation and its restricted time-dimension by the idea of continuous natural selection;
- In family matter, challenges to gender roles and relations.

There is more, and the process goes on: By now, even the idea of rigidly-fixed gender is gone!

And no short passage better captures the underlying conception of this new world-view than this one, from Max Born's 1954 Nobel Prize lecture in Stockholm:

> I believe that ideas such as absolute certitude, absolute exactness, final truth, etc. are figments of the imagination which should not be admissible in any field of science. On the other hand, any assertion of probability is either right or wrong from the standpoint of the theory on which it is based. This loosening of thinking (*Lockerung des Denkens*) seems to me to be the greatest blessing which modern science has given to us. For the belief in a single truth and in being the possessor thereof is the root cause of all evil in the world.

This has been a truly radical transformation, playing out across a century or more, affecting the way we see the world in science and art as well as the way we understand the human person, and following thereon, the implications of these changes for social

relations. But new forms do not render the old ones useless or unimportant: We can still derive immense pleasure and insight from the earlier phases of classical music and painting, and even Einstein's new cosmology does not abolish Newtonian physics. (This is not the case across the board, to be sure; in some areas, such as psychology and evolution, the new ways of thinking render the older ones obsolete.)

What the new forms do is to present a radical challenge to the limitations on thought and sensibility represented by the older ones, and to present a broader and deeper apprehension of ourselves and the world. And precisely because the constructions of modernity were so thoroughgoing and radical in their challenges to established forms, *the eventual reaction against them in Nazi ideology was equivalent in its radicalism both in breadth and depth.* This reaction did not limit itself to repealing what had happened in the previous half-century; on the contrary, it recreated an ideal of society and its leadership cadre that leapfrogged all the way back to medieval times!

This blueprint can be seen most clearly in the activities and plans for the future carried out during the war years by SS leaders under Himmler's authority. These plans and activities included:

1) resettling all of the Soviet Union up to the Urals (after exterminating some of the existing population and enslaving the rest) with landed agricultural estates owned by German "soldier-farmers," living in medieval-style houses, who would sow ancient grains and tend ancient cattle breeds;

2) creating a SS training academy in a castle at Wewelsburg, furnished with items including a shrine dedicated to the Holy Grail and artifacts from Roman and Bronze Age tribes;

3) spending vast funds on researching the origins of the "Aryan" race, including sending an expedition to Tibet.

The better-known abusive reactions of the regime to modernity included labeling relativity theory as "Jewish physics," proscribing most twentieth-century classical music and jazz, and creating a "museum of decadent art" (of course, this did not prevent kleptomaniacs among the regime's grandees, especially Göring, from stealing truckloads of valuable modern paintings from their Jewish owners).

The most salient fact about the hatred of modernity in Nazi ideology was its obvious impotence. Neither on an intellectual nor an aesthetic level did this opposition mount – or even really try to mount – any kind of meaningful response to what it condemned. What aesthetic responses existed were either banal, kitschy or backward-looking, some examples of which are the sculptures made for Hitler's chancellery and especially Speer's grandiose designs for massive buildings to be built in Berlin, Linz, and elsewhere. On an intellectual level the responses were even more pathetic, as Nazi ideologues tortured people, facts, iconography, and history in attempting to reinforce their bizarre racial and medical theories. No new scientific discoveries of any kind were made, partly because some genuine scientists such as Otto Hahn refused to continue his work, and even Heisenberg's attempt to build a working nuclear reactor using controlled fission – a necessary prelude to making an atomic bomb – was simply amateurish.

Thus, there was no effective or meaningful creative outlet for the Nazi rage against modernity in the realms of thought and art, nothing at all which could be pointed to as representing either a satisfying vindication of the campaign itself or a clear sign of its triumph. Only one avenue remained open for discharging the contents of that immense psychic reservoir of inarticulate hatred and suppressed self-loathing with which the Nazi mind was filled to overflowing. That avenue was the attempt to carry out the complete extermination of European Jewry. Hitler's rhetoric abounds with an indictment of the Jews as the source of all the ills of the modern age, but the bill of particulars was always an inchoate miscellany of bizarre lies that never even rose to the level of half-truths. To ask for more convincing evidence to support the case would have been, of course, beside the point. To be sure, the pervasive, ancient, latent, ever-stirring antisemitism rooted in Christianity had prepared the way, but this time was different. As Franz Werfel wrote – about other, similar events – in his 1933 novel, *The Forty Days of Musa Dagh*, his unwitting anticipation of the Holocaust: "The old sporadic fanaticism of religious hatred had been skillfully perverted into the cold, steady fanaticism of national hate."

From its beginnings, this trial was never about how convincing the evidence of guilt was. Rather, it was about generating a process of thoroughgoing social and institutional mobilization within German society that would be sufficient to allow the leadership to implement unopposed the *Endlösung* (the "final solution" to "the problem of the Jews"). The details of German history from 1933 to 1944 show how carefully and methodically Hitler and his henchmen undertook this mobilization. It proceeded in fits and starts, during the first phase known as the time of persecution, starting with the initial wave of repression soon after the accession to power, sending some Jews and others to the first concentration camps, then the racial law in 1935, *Kristallnacht* in 1938, and many other steps; in between those steps there were periods of quiescence, which lulled the eventual victims into false hopes. Even the minor forms of harassment had an utterly heartless character, such as the order banning Jews from owning household pets, leading to the spectacle of weeping people bringing their dogs, cats, and birds to places where they would be killed.

The second phase, the period of extermination, also developed in stages at first – but after the "Wannsee Conference" in early 1942 it was to accelerate rapidly. Earlier there had been much speculative discussion within the regime about other "solutions," notably the proposed mass deportation of European Jews to the island of Madagascar, where they would be allowed to starve to death, a plan which could not overcome the prospect of confronting the might of the British Navy. But a powerful accelerant had been added in the middle of 1941, namely the fact that huge numbers of Jews in eastern Poland, the Baltic States, Ukraine, and Russia had newly and suddenly fallen under Nazi control after the invasion of the Soviet Union, far more Jews than the total existing in all of Western Europe.

By early 1942 the institutional mobilization for extermination had been completed. The security service (SS and SD) had an iron grip on the population, both within Germany and throughout the occupied nations, and took the lead under the command of Reinhold Heydrich and Heinrich Himmler. The officers and soldiers of the *Wehrmacht* on the Eastern Front, including at the most senior levels of command, had been

conditioned to play their supporting role; the mobile extermination squads were already active in the East, carrying out mass shootings; the concentration camps for the exploitation of slave labor were expanding rapidly in number and size; and experiments were under way to determine the most efficient methods for mass murder on a scale never before contemplated in what was once a civilized nation. Even the euphemisms designed to disguise the enterprise had been devised.

When a sufficient level of mobilization had been attained, this countervailing, sinister *Schwerpunkt* came into being: A concentrated mass of overwhelming military and police power, supported by a broader administrative bureaucracy, devoted to achieving a single objective, one that had no precedent (so far as I know) in earlier human history – namely, the complete extermination of an entire people and its culture.

From the westernmost borders of Europe eastwards into the depths of Russia, from Norway's far north to the boot of Italy, a gigantic machinery, operating at a frantic pace, was set in motion, which spared no effort or cost and had substantially realized its objective within a mere three years. The centrality of this single objective to the regime's conception of itself is revealed in the last major phase of its operations. In June of 1944 the Allied landings in Normandy were matched on the Eastern Front by the Red Army's "Operation Bagration" (launched to the day on the third anniversary of the German invasion); thereafter the military conquest of Germany by these combined armies was only a matter of time. When the Red Army paused to resupply a few months later, it had destroyed once and for all Germany's Army Group Centre, its most powerful military formation, and had arrived at the eastern borders of the Reich and of Hungary. And what had the Germans been doing in the meantime, with their nation coming under mortal threat from its sworn enemies? During a mere two months after May 15 of that year almost half-a-million Hungarian Jews were rounded up and deported to the extermination camps in Poland and summarily gassed and burned.

The kind of historical dialectic outlined here – the way in which a radical movement of thought and sensibility, challenging centuries-old established traditions, elicited a horrifying and equally radical reaction – is not without historical precedent. For

example, monarchs seeking to uphold the Catholic Church's monopoly of "spiritual" power and authority against the challenge of Protestantism unleashed a bloodbath on the European continent during the Thirty Years' War in the seventeenth century. Yet the differences between these two are not trivial. The twentieth-century case occurred in the context of greatly enlarged human powers brought by modern science and technology and its supporting structures – bureaucratic and administrative organization, advances in mass communication and mass destruction, and the sheer geographical reach of conflict. The four years of bitter struggle between Nazi Germany and the Soviet Union was the largest and most destructive war in human history. And by no means did all of the horrors either contemplated or attempted by Hitler actually come to pass; the war had a number of decisive turning-points, the outcome of which was a close-run thing. It could have been even much worse than it was.

It was the terrible fate of the European Jews, whose impoverished and oppressed communities had given the world so many of the creative geniuses – out of all proportion to their numerical share of the larger population – who had helped to bring modernity into being, to suffer and die – again out of all proportion – without ever knowing why they had been singled out: Rounded up without warning and taken by the thousands to be shot on the edge of hastily-dug pits; or thrown out of their homes by the tens of thousands, stripped of their possessions, crammed into cattle-cars without food or water, families torn apart at the selection points at the camp's entrance amidst barking dogs and the screams of armed thugs, shorn of their hair, and shoved naked into sealed cellars, until choking on the deadly gas they were finally released from the terrors of their last days.

Paul Antschel (1920-1970), who wrote poetry under the name Paul Celan, was a member of the Romanian Jewish community whose parents died in a Nazi concentration camp. Following is the final stanza of his famous poem, "Death Tango" or "Death Fugue"; the poem was recited in its entirety in the German Bundestag in late 1988, on the fiftieth anniversary of *Kristallnacht* – the "Night of Broken Glass," which occurred on 9 November 1938:

Black milk of daybreak we drink you at night
we drink you at noon death is a master from Germany
we drink you at nightfall and morning we drink and drink
death is a master from Germany his eye is blue
he strikes you with leaden bullet he strikes you true
a man lives in the house your golden hair Margarete
he sets his dogs on us he gives us a grave in the air
he plays with the serpents and dreams death is a master from Germany

For those in the camps the savage torment they endured did not end for them even upon their death, as other prisoners (whose turn would come) yanked the teeth from their corpses in search of gold fillings.

The historical agony of modernity and the reaction to it is not finished. More recently radical Islam presents the same kind of *totalizing* rejection of the modern, not just new ways of thought and expression, but the entirety of the new form of social relations, above all the position of women in the family and public life. The same kind of backward-looking utopia motivates its adherents. In his congratulatory message to his followers after the 2001 attack on the World Trade Center, Osama bin Laden extolled the memory of "Al-Andalus," the great Muslim civilization in Spain which flourished from the eighth to the fifteenth centuries. The Islamic State seeks to revive the long tradition of the caliphate in the Arab world, where political and religious rule were completely merged. Wahhabism, the severely conservative form of Sunni Islam which Saudi Arabia's oil money has promoted throughout the Islamic world, reaches back through an eighteenth-century preacher to its roots in medieval theology. Their Shia brethren in Iran long for the return of the Mahdi, the twelfth Imam, last seen in the ninth century, whose second coming will usher in the hoped-for end of days.

And at the core of the violent political hatreds expressed by both extremist forms of Islamism – Arab and Persian, sworn enemies of each other – we find not only the values of modernity but the State of Israel. At times, it seems inevitable that this hatred must explode anew in a bloody conflict that consumes the entire region.

APPENDIX: A NAZI PHILOSOPHY OF DEATH

A book by the French scholar Emmanuel Faye contains the following passage (on page 305) from one of Martin Heidegger's four so-called "Bremen Lectures," written in 1949 and entitled "The Danger." (The original German of this passage, taken from volume 79, page 56 of Heidegger's *Gesamtausgabe*, is quoted in the footnote section of Faye's book on pages 406-7.)

> "Hundreds of thousands die *en masse.* Do they die? They perish. They are put down [*umgelegt*]. Do they die? They become supply pieces [*Bestandstücke*] for stock in the fabrication of corpses. Do they die? They are liquidated unnoticed in death camps. And also, without such -- millions in China sunken in poverty perish from hunger. But to die means to carry out death in its essence. To be able to die means to be able to carry out this resolution. We can only do this if our essence desired the essence of death. But in the middle of innumerable deaths the essence of death remains unrecognizable. Death is neither empty nothingness, nor just the passage from one state to another. *Death pertains to the* Dasein *of the man who appears out of the essence of being.* Thus it shelters the essence of being. Death is the loftiest shelter of the truth of being, the shelter that shelters within itself the hidden character of the essence of being and draws together the saving of its essence.
>
> "This is why man can die if and only if being itself appropriates the essence of man into the essence of being on the basis of the truth of its essence. *Death is the shelter of being in the poem of the world.* To have the capacity for death in its essence means to be able to die. Only those who can die are mortals in the apposite sense of the word."

The Nazi regime used the phrase *Lebensunwertes Leben* ("life unworthy of life") for those fit only for extermination. What can one say to this passage from Heidegger's postwar writings except: *Is it not fitting that those who while alive were deemed to be a form of "life unworthy of life" should, at the brutal termination of their existence in the extermination camps, be deemed to have suffered a death unworthy of death?* In the passage quoted above, the imprecision and allusiveness of the author's prose is put at the service of a train of thought that is shocking in its overt indifference to suffering and

injustice, but perhaps less shocking in the case of an unrepentant servant of Nazism who had willingly configured his own philosophy to make it accord with that evil regime's basic tenets.

Section Two: Pathways to Utopia

Chapter 5: A Utopia for our Times

> The projections and results presented in several peer-reviewed publications provide evidence to support a physically plausible Global Mean Sea Level [GMSL] rise in the range of 2.0 to 2.7 meters [by 2100], and recent results regarding Antarctic ice-sheet instability indicate that such outcomes may be more likely than previously thought…. Though the GMSL rise scenarios are primarily framed for overall changes occurring by 2100, it is important to recognize that GMSL rise will not stop at 2100; rather, it will continue to rise for centuries afterwards.
>
> NOAA (2017)

> Some scientists point out that during the last ice age, ice sheets similar to West Antarctica's formed in other ocean basins. But as the ice age ended and the oceans warmed, all of them collapsed. These experts have started to think that West Antarctica, as a fragile holdover, is basically a disaster waiting to happen — [a collapse of] the most vulnerable parts of the West Antarctic ice sheet could raise sea level by 10 to 15 feet… "We could have a substantial retreat on a time scale of 10 years," said Robert A. Bindschadler, a retired NASA climate scientist who spent decades working in Antarctica. "It would not surprise me at all."
>
> Justin Gillis, "Antarctic Dispatches" (2017)

Platforms for Utopia

Introduction.

UTOPIA IS A WORD FROM THE GREEK language meaning "no place" or "nowhere," but which is also a *double-entendre*, since it can, as *eu-topia*, mean "good place," and sometimes "ideal place." The tradition of modern utopian thought begins with *Utopia* (1516) by Sir Thomas More (1478-1536) – Saint Thomas More within the Catholic Church – lawyer and Lord Chancellor of England, and a humanist scholar schooled in Latin and Greek. He was beheaded by Henry VIII for his refusal to sanction Henry's divorce from Catherine of Aragon and to recognize his monarch as head of the Church of England: a martyr to his Catholic religious faith, perhaps the only famous churchman ever who was also praised by the officially atheistic regime in the Soviet Union – in its case, for the "communist" structure of his ideal society.

Thomas More's *Utopia,* as well as Tommaso Campanella's *City of the Sun* (1602) – which shares More's communism – and Johann Andreae's *Christianopolis* (1610), steeped in

religious mysticism, were strongly influenced by the first work of its kind in European civilization, Plato's *Republic*. But it is only in Francis Bacon's unfinished short work, *New Atlantis* (1627), that the authentically modern form emerges, because it was Bacon who introduced the commitment to science and material progress into the utopian vision of the future.

The first stage of industrialism in the early nineteenth century breathed new life into this genre; many authors responded to it by combining utopianism, industry, and socialism. The most important early writers and crusaders for this vision were Henri de Saint-Simon (1760-1825), Charles Fourier (1772-1837), and Robert Owen (1771-1858). Owen was also an industrialist who tried to put his ideas of a "new moral order" into practice in the cotton mills at New Lanark. (Throughout the nineteenth century various types of experimental "socialist" communities sprang up, especially in the United States.) The four notable fictional works on the English-language side of this tradition appeared towards the end of the century: Samuel Butler's *Erewhon* (1872), Edward Bellamy's *Looking Backward* (1888), William Morris's *News from Nowhere* (1890), and William Dean Howells's *A Traveler from Altruria* (1894).

Morris's utopia is especially notable because he is one of the first writers in this genre to turn away from the belief that industrialism is an appropriate economic basis for a harmonious society of the future. However, what emerged as the dominant tradition in this imaginative fiction thereafter placed a strong emphasis on new gadgetry made possible by modern technology. The fifty-four novels by Jules Verne (1828-1905) provided enormously popular stories along these lines, as did many of the fifty novels written by H. G. Wells (1866-1946). Wells is sometimes called the "father" of science fiction, although this can be misleading, since many of his novels, especially the later ones, emphasized social rather than technological themes.

In the twentieth century, this tradition split into two streams. First, there is the literary stream which may be called "futuristic fiction." Here important novelists turned utopia into "dystopia," a bleak vision of possible futures. Its first great expression occurs in *We,* written in Russian by the naval engineer Yevgeny Zamyatin in 1920, but published first in English translation in 1924. The better-known works that followed were Aldous Huxley's *Brave New World* (1932) and George Orwell's *1984* (1948). There is also the trilogy by C. S. Lewis, *Out of the Silent Planet* (1938), *Perelandra* (1943), and *That Hideous Strength* (1945), the last being a rather dismal account of a domineering brain in a vat. Many women writers, notably Doris Lessing, Ursula Le Guin, and Margaret Atwood have entered this field; Atwood's work includes *The Handmaid's Tale* (1986) and her trilogy, *Oryx and Crake* (2003), *The Year of the Flood* (2009), and *Maddaddam* (2013). In this stream, the story line does not depend primarily or even importantly on machines or gadgets, especially those which supposedly appear in the future. Rather, the focus is on forms of social organization that are presented as being a possible outcome of present-day trends.

This brief sketch shows that the genre of futuristic fiction attracts two different types of authors. One is the writer of "serious" prose about social trends, who chooses the novelistic form as a way of dramatizing the account and, possibly, appealing to a wider audience. Such authors commonly produce no other kinds of fiction. The other type is

the accomplished literary figure, such as Huxley or Atwood, who writes novels, short stories, and poetry, and who occasionally chooses a futurist setting.

Second, there is the science fiction stream, dominated by the mass-market paperbacks and Hollywood horror films that emphasize futuristic gadgetry. Yet there are important works in this sub-genre which rival those in the literary stream in terms of imaginative power and which focus more on social as opposed to technological issues. Notable here are the remarkable series of books by Philip K. Dick (1928-1982), including the three stories that have become major movies (*Blade Runner, Total Recall,* and *Minority Report*); by the Polish writer Stanislaw Lem (1921-2006), author of *Solaris, Mortal Engines, The Cyberiad,* and many others; and by Isaac Asimov (1920-1992), especially *I, Robot* and the three novels of his *Foundation* series. There are many other writers, such as John Brunner, especially his *The Sheep Look Up* (1972), who have produced interesting novels of this type.

Four Contemporary Platforms

1. **The Amish/Hutterite Platform**:

This model encompasses some traditional, religiously-based communities with a very long history, and excludes the short-lived and often fatal experiments led by delusional zealots. The Amish, Hutterite, and Mennonite communities, which exemplify this social formation, are all branches of the Anabaptist form of Protestantism – those objecting to infant baptism – which arose in the early sixteenth century. The fate of the Hutterites was typical of these peoples. Named for their early leader Jacob Hutter (1500-1536), who was severely tortured and then burned at the stake, they were triply hated for their religion, their pacifism, and their devotion to a communal life. Forced to flee eastwards out of Austria and Moravia to escape persecution, the Hutterites eventually settled as far as east as Ukraine before turning west again and migrating to the United States, Mexico and Canada in the 19th and 20th centuries. The Amish (who separated from the Mennonites) and the Mennonites emigrated earlier to North America, beginning in the 18th century.

The Amish and Hutterites were more insistent than were the Mennonites on separation from the larger surrounding society, and their communities exist exclusively in rural farming areas. Hutterites differ from the Amish in that most of the property they own is held in common; together they number fewer than 200,000, whereas the Mennonites, most of whom do not live apart from the larger society, number in the millions. Hutterites appear to be more open to modern technology than do the Amish, most of whom are Old Order Amish who limit or prohibit access to automobiles, telephones, and power-line electricity. The distinctive, separate communities of rural Amish and Hutterites have endured with their own unique lifestyles for almost 500 years.

2. **The Endlessly-advancing Technology Platform**:

This model owes its beginnings to *New Atlantis*, a remarkable, unfinished utopian fantasy, less than 50 pages in length, written by Sir Francis Bacon (1561-1626) during the last years of his life and first published in 1627. It tells the story of "Solomon's House," a private scientific research foundation which is the dominant institution in a society located on a small, isolated island. Its motto is given as follows: "The end of our foundation is the knowledge of causes, and secret motions of things; and the enlarging of the bounds of human empire, to the effecting of all things possible." Driven by a single grand idea which animates it, within the small genre of utopian fiction it is doubtless the most perspicacious anticipation of a society driven by its total commitment to fostering scientific and technological progress – in other words, the one which came to fruition in the West over the course of the 20th century.

On this platform humanity is dedicated to becoming the "masters and possessors" of nature, *without envisioning that humanity itself should change in any fundamental way*. In other words, we will always know who "we" are, no matter what far-ranging transformations in earthly existence – our lifestyles and the satisfaction of our needs – we manage to achieve for ourselves whilst always searching for more and better means of gratification.

I have been so enamored of this perspicacious little book that my sisters and I adopted the basic premise and operational rules of our Yucca Settlement enterprise from it. We designed a privately-owned scientific establishment, to be the first of many like it around the world, that would be completely free of any external control and of any dependence on a national government for funding, and with this freedom we resolved to exercise an ethically-based oversight of the consequences of scientific and technological applications. Francis Bacon was the first to describe such a plan:

> The end of our foundation is the knowledge of causes, and the secret motions of things; and the enlarging of the bounds of human empire, to the effecting of all things possible…. And this we do also: we have consultations, which of the inventions and experiences which we have discovered shall be published, and which not; and take all an oath of secrecy for the concealing of those we think fit to keep secret: although some of those we do reveal sometime to the state, and some not.

Although unlike in Bacon's original fantasy we do not live on an isolated island, we are sufficiently resourceful to defend our independence against all comers, and the stability we provide in our defined territory is an advantage that is appreciated by the residue of national governments on our continent, which leave us to our own devices. Finally, we dedicated our enterprise to the larger good of human welfare, as Francis Bacon himself did.

3. **The Hyper-Technological Post-Human Platform:**

Here everything is up for grabs—the platform included. Within this conception, the meaning of existence is to be self-created continuously, primarily with the assistance of our technological mastery of matter and energy. Or rather, it might actually be said that there is no "meaning" *per se* for us humans, since it is identical with our activity itself and thus entirely open-ended, including the possible transition to another "nature" entirely (as engineered by us). In its most extreme versions this fantasy is now synonymous with our willingness to pursue the relentless advance of artificial superintelligence to the point where humans are replaced entirely by their more talented creations.

The strange, eerily calm matter-of-factness one finds in the language used by all the devotees of a post-human platform is a clue to its origins in the older idea of a technological imperative: "It's going to happen, be sensible, resistance is futile, don't waste your time and energy trying to fight it." In the notion that we should consecrate our biological mind and body to becoming a simple platform for a machine superintelligence, we have reached the final stage in celebrating a meek submission to necessity and fate. The apparent triviality and lack of manifest heroism in their labor, namely the writing of computer code, appears to have nothing overtly in common with those German soldiers in the 3rd SS Panzer Division who, in the nineteen-forties, strutted around Russia wearing the *Totenkopf* (death's-head) insignia on their uniforms, but make no mistake, this more recent obsession is also a cult of death.

One is reminded of the powerful metaphors of death which accompany Melville's picture of the female labor-force, whose life is literally sucked out of them by the

machinery they serve in a New England paper mill, in his extraordinary 1855 short story, "The Paradise of Bachelors and the Tartarus of Maids":

> Something of awe now stole over me, as I gazed upon this inflexible iron animal. Always, more or less, machinery of this ponderous, elaborate sort strikes, in some moods, strange dread into the human heart, as some living, panting Behemoth might. But what made the thing I saw so specially terrible to me was the metallic necessity, the unbudging fatality which governed it.

Perhaps the best epigraph for this platform would be the following sentence from Nietzsche's *Beyond Good and Evil*: "And when you stare persistently into an abyss, the abyss also stares into you." The abyss captures our imagination and freezes it, so that we would never be able to see the ominous downside coming right at us. In fact, this is an apt characterization because there is *nothing* to be seen when gazing into such an abyss: Neither the presumed need to do something so bizarre as to abandon humanity's past, nor the completely mysterious benefits to be gained thereby, nor the utterly unknown ultimate consequences, nor even a decent summing-up of what will be lost in this transformation, are at all evident. Like the case of the Abrahamic religions, this fantasy is a testament to how the idea of absolute power unhinges the mind.

4. **The** "**Art Nouveau" Platform**:

William Morris (1834-1896), an English artist, textile designer, writer, and speaker, published his *News from Nowhere* in 1890, the last significant work in the modern tradition of utopian thought. The defining characteristic of his vision was its emphasis on the satisfactions derived from craft and farm labor, assisted by mechanical aids where necessary, taking place in a largely pastoral and small-town setting. The outputs of craft labor are described on terms of his strong aesthetic sensibility. His book sets its face against almost everything in the modern world already evident in his day – large-scale machinery and industry, ever-advancing technologies, the factory system with its specialized, repetitive tasks, huge cities and their slums, and working-class poverty.

Morris thought that if the ordinary worker, male and female, were not faced with the need to accept exhausting, repetitive work in factories, or similarly exhausting labor in farm fields, simply to earn a living at the edge of poverty, but were promised by a more just social order a comfortable yet modest standard of living, then the practice of, and pride in, skilled craft labor could offer the deepest sense of life-satisfaction.

The first two of these platforms have actually existed in a number of variations; the latter two have remained unrealized until now. All of them share with the rest of the utopian promise the basic concept that there is only one preferred type of ideal social organization for all of humanity, or at least only one that really matters. *This was a serious mistake, in that it tried to squeeze all of the beneficial diversity of human experience into a single too-small mold*. Apart from the one I have called the post-human platform, which to me represents a deranged state of the human imagination, all of the other three have valuable elements which ought to be preserved, further developed, and – most importantly – tightly associated within a larger whole.

By this I mean that the main aspects of all three can coexist within a larger totality that is at the same time segmented, so that a common interest binds all three together without erasing the equally important differences they represent. Since all three have aspects of great human value worthy of preservation and indeed enhancement, the larger association of which they form the parts ought to be dedicated to the protection of the well-being of all of them, whilst refraining from interfering with the large measure of autonomy which each deserves to have.

A Mixed Platform for Utopia in our Times

The "Mother Settlement" we established at Yucca Mountain some decades ago, which has since been emulated with variations by like-minded groups on other continents, consciously strove to avoid the errors in the many earlier and too-limited utopian designs. We sought to mix the advantages of three platforms: first, using advanced technologies and light industry, operated by computerized machinery and robots; second, a strong craft-labor and small-workshop emphasis in all human settlements; and third, relying on a network of Amish and Hutterite farming communities for most of our food supply. The single unique characteristic of the Yucca Settlement consists in the genetically-modified members of our First and Second Generations, whose special sense of empathy was engineered to be heritable, according to a protocol designed by my parents – as described in the first volume of this trilogy, the book *Hera, or Empathy*. This intervention will not be repeated in the future, whereupon the modification will pass into the human genome and take its chances in our evolutionary future.

At the moment, there are a total of ten fully-operating settlements in all, located in North and South America, Europe, Australia and New Zealand, and Singapore, and preliminary steps are under way for constructing additional bases elsewhere. They vary in total size from a few hundreds of square kilometers to – in our case – some hundreds of thousands of square kilometers, all inside strongly-protected perimeters. All of them exist amongst the tide of social unrest outside their secure borders that has proliferated in response to sea-level rise, serious regional warfare, and endemic terrorism.

Figure 7 Looking toward Black Cone across Crater Flat from the crest of Yucca Mountain (Photo: W. Leiss)

This unrest has been worst in Asia, since enormous populations already lived at the prevailing sea-levels before the oceans began to rise, and where in many regions there are no longer functioning national economies or governments. The continent of Africa has been ravaged by the severity of climate-change effects near the equator. Most of Western Europe has been stabilized, but with sharply lowered economic productivity, in places having been reduced to pre-industrial levels, and weakened

political structures. The economies in Central and South America have been ruined everywhere, leading to endemic mass population movements against the southernmost areas of the United States and the quick destruction of its border wall. The Eastern seaboard of the United States has been largely emptied due to the rising ocean and severe storms, and although the national government and its military arm still functions, the country's economy and political structures have been badly weakened. Needless to say, there is no functioning international machinery of governance or assistance.

Since the Yucca Settlement is the oldest of its kind, we have had the most time during which to design and develop our internal social structure and economy. The high-technology operations function within a sealed inner perimeter, staffed by skilled technicians who have been recruited from among our own people and from outside, since we have more applicants for admission than we can accept from those fleeing the chaos around us. These operations are made up of our power supply (all electric, from solar, wind, and small nuclear units); our Machine Intelligence Unit; our electronic and metallurgical factories, with entirely automated production lines using computer-controlled robotic machinery (recycling metals from the vast supplies of junk available outside our borders); our large, intelligent, mobile robot "army," for assistance with tasks requiring motility; and our R & D facilities. All of this infrastructure has been placed between Yucca Mountain and the city of Las Vegas; the city hosts our large university, technical institutes, laboratories, medical facilities, staff apartments, and administrative offices. On the outskirts of the former city is the old Nellis Air Force Base, where our military and security units are based.

Supporting human villages are scattered throughout the larger territory contained within our external perimeter, stretching eastwards from southwestern Nevada to the California coast. They are made up of small dwellings, schools, libraries, auditoriums, playgrounds and athletic facilities, warehouses, greenhouses, gardens, vineyards, studios, clinics, farms, barns, a few administrative offices, shops where goods are exchanged, craft breweries and distilleries, and large numbers of small workshops. Transactions are carried out by barter and gift exchange, with ration cards used only sparingly; there is no currency.

Career options consist largely of two paths, one for training in skilled craft labor, the other for streaming into higher education for professional, research, and technical roles at the university, labs and institutes, and oversight of factory production. All young adults serve a two-year training period in either the police or military services and remain attached to a militia unit thereafter. The social ethic of the Yucca Settlement demands that the residents of these villages, situated outside the inner perimeter where the staff of the high-technology sector operates, are entitled to a standard of living and all essential services equivalent to what is enjoyed by that staff and their families.

For these villages, there is an Intranet for personal communications among all its residents, but connections to the World Wide Web are available only in schools and libraries, where Internet traffic is carefully monitored. Some highly-specialized goods are exchanged by ship traffic among the group of scattered settlements, which are also connected via satellite communications with each other. Teams of researchers and medical and technical specialists regularly are posted from one settlement to another for varying periods of time, for advanced training or assistance. All of the settlements seek to increase their populations and secure territory by admitting as many refugees as possible and by recruiting persons with special skills from the surrounding areas. The nearest governments are encouraged to treat the settlements as sovereign territories, and settlements are admitted to the official group of the same only when they have the requisite military and economic means to enforce the security of their perimeters.

Finally, the Yucca Settlement is known among its peers for, among other things, its long-standing and harmonious relationship with its Amish and Hutterite communities. In the western section of its territory in Southern California, most of the land is given over to the large farming operations run by these communities. In return for our guarantee of protection from malcontents and terrorists, they provide us with the huge agricultural surplus produced by their farms to feed our population. With respect to their own customs, they are allowed full autonomy to manage their internal affairs, and freedom from interference in those affairs from the other residents in our human settlements. They may, however, call upon us for assistance in police matters, and of course our medical facilities are open to them at all times.

A Measure of Success

This hybrid utopia just described, if it unfolds as it ought, has a chance of becoming a good society – neither perfect nor ideal, but a decent and enduring basis for human well-being. We expect that it will be tested severely by the many social and environmental challenges that will prevail by the end of the twenty-first century. Once these challenges work their way through the existing societies, we think, there will be little or nothing left of the large-scale industrial economies or political structures in place at the beginning of that century.

First, the new social formation will be tested against persistent stresses resulting from vast population movements happening around the globe, millions on the move, desperate to find refuge from the rising seas, the hotter temperatures, and the violent storms induced by climate change. These stresses will be compounded by the accompanying regional wars and endemic violence, enhanced by the occasional use of nuclear, chemical, and biological warfare agents, as well as by natural plagues and pandemics. Only by being able to defend their secure external perimeters against those

unfortunates whom it cannot take in and shelter will the representatives of this new social formation be able to survive in the midst of such long-running chaos.

Second, this hybrid utopia must be able to create an economy that is sustainable over the long run, and one which provides an adequate, but not opulent, standard of living for all of its citizenry. It must be able to adapt to the changing climate as well as to the social stresses already described. Also, the young people growing up in the villages must find opportunities for moving into the staffing positions in the high-technology sector equal to those who grow up among the families of those already employed there. This will be a severe internal test, to see whether the new society can overcome the entrenched human tendency to skirt around equality of opportunity and reward inherited advantages. Should it be unable to do so, it is likely to fail to find and nurture the necessary talent for resisting the pressures from outside, and to collapse relatively quickly as a result. Finally, it must have the wisdom to nurture and respect the relationship between the inhabitants of these two sectors and their neighbors in the Amish and Hutterite communities.

Since we have embarked only recently on this new path, in the midst of severe social chaos and environmental change around the globe, my sisters and I continue to debate among ourselves how we should respond to these challenges as they unfold in the last quarter of the twenty-first century. The final chapter in Part Two presents a brief summary of these debates.

Should the solutions we have designed prove to be robust, modernity will have finally found an adequate response to the most serious challenge of all, namely, sustaining the scientific and technological enterprise – its defining historical achievement – without falling into the snares of either the twentieth century's desperate and destructive use of its immense powers, or the twenty-first century's grotesque fantasy of a post-human future.

Chapter 6: The Threat of Superintelligence

THE IDEA OF A "SUPERINTELLIGENCE" – or of various types of superintelligent entities – is best-known from the 2014 book of that title by Nick Bostrom. It is defined by Bostrom as an entity that has mental powers far beyond those of any human being in terms of general intellectual skills, especially in cognition, memory, and recall, and possibly even in social and behavioral skills. This idea was originally conceived as long ago as 1956, at the Dartmouth Artificial Intelligence Conference, with the contention that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."

The special value of Bostrom's approach is his detailed focus on risks, which is what sets it apart from almost all earlier work, where the term "singularity" was often used. A singularity in this context denoted the idea of a moment in time when machines became more intelligent than humans, whatever it is that one means by "intelligence" – the proponents of this idea often were content to be vague on this rather important topic. In older usage in the field of cosmology, the initial singularity refers to the state of the universe at it existed in the moment before the Big Bang; thus this later usage is not really in the same league, as momentous events go, and should more properly be called the machine singularity. Moreover, any reflections on the ultimate consequences of this event were for the most part cursory and superficial in the extreme.

For example, here is a passage from a long 2010 article, "The Singularity," written by David Chalmers: "*If there is a singularity, it will be one of the most important events in the history of the planet. An intelligence explosion has enormous potential benefits: a cure for all known diseases, an end to poverty, extraordinary scientific advances, and much more. It also has enormous potential dangers: an end to the human race, an arms race of warring machines, the power to destroy the planet.*" Let's pause for a moment and ponder the words we just read. The list of named benefits comes with the adjective "potential," a common synonym of which is "possible." In what sense – or under the terms of what bizarre type of calculus – would the simple *possibility* of achieving those named benefits justify taking the named risks – which include the rather untrivial possibilities of "an end to the human race" and the destruction of our planet?

"Potential" also means "probable," and neither of those terms, of course, denotes certainty, or, necessarily, anything close to certainty. If one has a good enough decision-analytic system in which probabilities are quantifiable, one can put some numbers on them – always providing that one also tries to quantify the uncertainty bands around the probability estimates. Even if one winds up with a low-probability estimate, one

should not forget to watch out for low-probability, high-consequence events, especially those in the zone of catastrophic outcomes.

> *So, for example, with respect to the quote from Chalmers above, if one were faced with the prospect that there were roughly equal probabilities of realizing either or even both the list of named benefits and the list of named adverse outcomes, who in his or her right mind would say: "That seems OK, let's proceed." If, on the other hand, the proponents were to contend that realizing the benefits is more likely than the opposite, would one not ask: "*How much *more likely?*
>
> *And if your interlocutors were to reply, "Well, no one really knows, since it's all so new, we'll just have to try it out and see what happens," my advice would be: Show them the door.*
>
> *At which point some of the advocates, being ushered on their way out the door, will look you in the eye and say, perhaps with a pained expression on their faces: "Unfortunately, the coming of the singularity is inevitable, resistance is futile, so make the best of it." They would bid you farewell and rush back to their labs, turning their attention to their favorite game, namely, forecasting how soon – in decades – the singularity would be upon us, like it or not.*

These twin themes of uncertainty in the calculation of future benefits and adversities, on the one hand, and a sense of the technological imperative which offers us no choice in the matter, on the other, are very old refrains in the reaction to the capitalism, industrialization and the machine age. The German sociologist Max Weber first called attention to the internally self-expanding character of this nexus, because the visible benefits derived from the process of "rationalization" (means-ends rationality) lead to calls for more of the same. But it has been clear to many commentators, since the middle of the nineteenth century, that various upside as well as downside prospects in the new reality were closely linked; most observers, however, expressed confidence that, at least in the long run, there would be a huge surplus of net benefits, for society considered as a whole, but of course not necessarily for every person or group.

This was at least a plausible contention – until the arrival of the atomic and hydrogen bombs. By around the first half of the nineteen-fifties, both of the nuclear-armed superpowers, the Soviet Union and the United States, then mutually locked in to the

Cold War, had vast stockpiles of hydrogen bombs, orders of magnitude more destructive than the simple bombs dropped on Hiroshima and Nagasaki, as well as the means to deliver them to all the major cities in the territories of their enemies. From that point onwards, all of the most advanced industrial societies faced the common prospect of annihilating destruction. And the likely onset of secondary consequences, as a result of "nuclear winter," would spread those horrors more widely around the globe. Whether *any portion* of the heritage of human civilization accumulated until that time would have survived this madness was an open question.

This was the point when the benefit-risk calculations associated with ever-advancing industrial technologies collapsed, although most people would not have sensed it at the time. At this point the species *homo sapiens* as a whole faced an *existential risk* perhaps for the first time since the onset of the last Ice Age, some 110,000 years ago (assuming the origins of the species *homo sapiens* to be about 300,000 years in the past). By the term "existential risk" I mean a situation in which there was posed the very real question about whether there was any remaining possibility of living a truly human life in the future. This was the first time this question could sensibly be posed since the beginning of the Industrial Revolution. And if it could plausibly be argued that this existential risk was a "logical" outcome of that Revolution, in which immensely powerful new technologies had developed in the absence of effective controls over their deleterious use by nation-states, then that argument could not avoid entertaining at least the possibility that the whole course of its history had been a mistake.

Since the worst had not happened at the time when tensions between the Soviet Union and the United States were at their height, most people could be forgiven for thinking that the risk was a remote one, but if so they would have been wrong. Not only during the 1962 Cuban Missile Crisis, but during at least two or three other episodes, occurring as late as September 1983, when automated surveillance routines in both countries mistakenly detected incoming nuclear missile strikes from the other side, there were narrowly-averted catastrophes. After the collapse of the Soviet Union in 1991, both Russia and the United States retained large arsenals of nuclear-tipped missiles, ready in their silos to be launched at a moment's notice, and other nuclear powers remained ready to initiate regional nuclear wars.

Those who paid some attention to the nuclear threat thus knew full well that humanity had been exposed to truly catastrophic risk scenarios since the 1950s. This older existential risk had still been hanging over our heads at the time when newer ones appeared, one of which was represented by the idea of the machine singularity. And during the time when this idea was first being fleshed out, some significant neuroscience research – using monkeys and rats – was being designed under the acronyms MBI (machine-brain interface) and CBI (computer-brain interface). It was inspired by the need to develop solutions for people living with severe forms of paralysis.

First implanting hundreds of thin microfilaments into the somatosensory and motor cortex regions of a monkey's brain, the researchers recorded the electrical signals emitted by neurons as the monkey watched movement toward a desired goal (a bowl of grapes!). Then they developed a computer program capable of translating that cascade

of signals into digital motor commands; eventually, by precisely synchronizing the monkey's brain signals with the computer's machine-language simulation, the researchers were able to train the monkey to first, move a set of robotic arms on an avatar on a computer screen by waving its own arms, and second, to *move the robotic arms on the screen by just thinking about doing so*. (This capability is due to the fact that the brain's mirror neurons in the premotor cortex are activated both when an action is performed and when the same action performed by another individual is observed.)

A spectacular demonstration of this effect was carried out in 2008, when a monkey in a lab at Duke University in North Carolina controlled the walking action of a robot on a treadmill in Japan just by thinking about it. Later research under the acronym BTBI (brain-to-brain interfaces) linked the brains of three monkeys using implanted electrodes and succeeded in having them jointly control a robotic arm – finding that the arm was most effectively controlled when the three monkeys collaborated on the task.

Figure 8 A Rhesus macaque in Kinnerasani Wildlife Sanctuary, Andhra Pradesh, India

During some remarks made in 2014, the Duke neuroscientist who pioneered this type of research, Miguel Nicolelis, threw cold water on an idea, popular in the machine

singularity crowd, that we would soon be uploading vast digital information packets into our brains and, better yet, downloading the contents of our brains into digital storage media:

> Apart from tasks such as motor control for which BMIs can become very useful, mimicking higher-order brain functions, such as knowledge acquisition, memory storage, performance of cognitive tasks, and even consciousness, may be beyond the reach of binary logic, the basis from which all digital computers operate, no matter how simple or elaborate. An interesting corollary of this view is that we need not worry about the forecast that, in the near future, a "really smart" digital computer/machine will supplant human nature or intelligence. In all likelihood, this day will never come because, in a more-than-convenient arrangement, our most intimate neural riddles seem to have been properly copyright-protected by the very evolutionary history that generated our brains, as well as the very complex emergent properties that make it tick. As such, neither evolution nor neurobiological complexity can be effectively simulated by digital computers and their limited logic.

Not surprisingly, this considered opinion by someone who actually knows how the brain works had no noticeable impact on the devotees of superintelligence.

What is Superintelligence?

To his credit, David Chalmers did conclude his 2010 long article on this subject by mentioning in passing that "we" should "negotiate the singularity" – note the presumption of inevitability here – "by building appropriate values into machines." (Are "we" in charge of the process, and if so, who are "we"?) Sounds good. In earlier sections in his article he had speculated whether machines could become conscious entities, assuming we had earlier figured out what "consciousness" is, and whether, in the process of "uploading" human minds into digitized storage space, those minds would retain the self-conscious awareness they had possessed in their natural physical state. If a bit of humor might be permitted here, the calm matter-of-fact tone in which these possibilities are considered almost certainly would vanish in an instant were they

to actually occur: One can imagine, among the lab staff, a reprise of the famous movie line from *Frankenstein* (1931), "It's *alive*!" Followed by a race for the exits.

But like most other commentators on these themes at that time, Chalmers had not pursued sufficiently the implications of these particular issues. In particular, whether some kind of will, autonomous and self-generating, could appear alongside a machine-like consciousness? And if so, whether such a will could be of the pathological sort, as some human wills clearly are, and wish to inflict serious harms on others? And also, in such cases, whether such pathological wills would also be adept at *deception*, as psychopaths are known to be, in other words, at concealing their wishes from others in their environment until they had acquired the capability of realizing their malevolent intentions?

These supplementary issues, of rather significant import, were tabled, decisively, incisively, and inescapably, by Nick Bostrom in his 2014 book.

The idea of superintelligence is broader than just machine-based simulation: Bostrom suggests that there are a number of different ways in which such an entity might be brought into being by human action, including:

1. Artificial (machine) computer-based intelligence (AI) operating at a very high level of performance, based on algorithms utilizing neural nets and "deep learning," the circuitry of which runs at close to the speed of light, and thus orders of magnitude faster than the biological circuitry of the human brain.

2. "Whole-brain emulation," that is, scanning and thus digitizing an entire human brain, encompassing the entirety of its estimated 100 billion (10^{10}) neurons and all of their one quadrillion (10^{15}) synaptic connections, and then manipulating the result to increase this artificial brain's cognitive and other powers;

3. Selective breeding of humans who have been genetically-engineered to have superior brain functions. Obviously, this could only be carried out across a number of generations and lifetimes.

Networking many individual entities of the first two types above could also produce levels of performance that are orders of magnitude higher than any single entity operating alone.

These discussions immediately raise the key issue, which is known as the "control problem": If any entity exhibiting superintelligence were to come into being through human design, would it be strictly subordinate to human wishes and commands, for all time to come? Or, alternatively, could it in effect "escape" and develop an autonomous will and mode of action? Would it also thereby become "self-conscious," that is, aware

of itself and of its powers? More to the point, could it "escape" from human oversight and control by deceiving its human minders about what it was up to – pursuing what Bostrom calls "clandestine goals" – until it was powerful enough to resist their attempts to bring it back under human control? If so, would it then "wish" to dominate us or perhaps even exterminate us?

Excursus on Deception

Would it be foolishly hyperbolical to observe in this context that deception, duplicity, and dissimulation lie at the very core of being human? Is there *anything* in the human behavioral repertoire that reveals more about ourselves? How on earth would we entertain ourselves without the literature of deception and the great operas built on it – without *The Marriage of Figaro*, *Don Giovanni*, *Othello*, *Tosca*, and all the rest? Or without the relaxational fiction of spy stories and crime dramas, which would simply not exist as a genre in the absence of deceit? In the best of them – such as John Le Carré's *Tinker, Tailor, Soldier, Spy* – there are multiple layers of deception which must be methodically peeled away, one at a time, to reveal the truth hidden underneath.

And then, of course, there is the "real world" of love and marriage, corrupted by betrayal; fraud of every imaginable kind in financial matters and political life; lies everywhere in every corner of everyday life, as documented by research studies. Computer viruses and malware flood the Internet. Success in warfare is unthinkable without artful deception, without camouflage, decoys, the feint, misdirection, and *real* spy craft. In organized religion, do not the devout deceive their priests at every turn as to their blamelessness, and do not the churches hunt down heretics and those only pretending to believe in the one truth? Are courts not flooded with pleas of innocence from the manifestly guilty? Who among us would swear on pain of eternal damnation that he or she has never been deceitful?

Then, of course, there are the manifold forms of self-deception, wherein we willingly, gladly indulge ourselves in half-truths, fables, and fake news. But maybe we are exempt from personal responsibility for all this, if in fact we live inside a wholly simulated reality

– *The Matrix*? Finally, is *everything that exists* a gigantic fraud: Is the three-dimensional-plus-time universe we think we inhabit just an immense holographic projection of something else entirely, as certain quite reputable physicists suggest?

Clearly, deception in human affairs is no trivial matter. In opera, in the best spy craft, and in the more elaborate Ponzi schemes, it rises to the level of high art and grand passion. Used well in military operations, it can change the course of world history. The Allied Powers mounted successful campaigns to deceive Germany both about the invasion of Sicily and the site of the D-Day invasion on the French coast. In all three of the decisive battles on the Eastern Front during World War II – at the gates of Moscow in 1941, at Stalingrad in 1942, and at Kursk in 1943 – the Red Army used many complex strategic deceptions that were instrumental in bringing about the ultimate military defeat of the malignant Nazi empire; the Russians have a special word for it: *maskirovka.* The idea that a superintelligent machine might try to deceive us so that it can process all matter in the universe into paper clips seems, by comparison, to be somehow just pathetic. Or childish – innocent rather than malevolent.

On the other hand, just standing by with a bemused look on our faces as our earth's totality of both organic and inorganic matter is consumed by the first phase of the cosmological paper-clip caper doesn't seem like a very good response, either. In its evolutionary course to date humanity does not appear to have made much progress in constraining or abolishing the propensity to deceive. How likely is it, then, that if the same propensity should emerge in our superintelligent machines, anyone will be able to spot it and nip it in the bud? True believers think that, at some point in its internally-generated evolution, machine intelligence will reach a "takeoff" point, where recursive loops will further augment its capabilities at a speed impossible for us to realize what is happening, and thus impossible for us to monitor and oversee and, more to the point, to intervene.

Successful deception, both on the personal as well as the world-historical levels, is an exquisitely fine-tuned affair. In the most successful cases on a smaller scale, the victim never discovers the plot, even after it has run its course. In other cases, the damage has

been done long before the deceit is uncovered (think of the famous British espionage caper involving Philby, Burgess, Maclean, and Blunt). In the larger operations, where the *maskirovka* is busily concealing preparations for a lethal counter-punch, the timing of the unveiling is a critical aspect of the response. In all three of those fateful World War II battles on the Eastern Front (Moscow, Stalingrad, and Kursk), the Red Army played a dangerous and courageous game, allowing the enemy to advance perilously close to its objective, sucking them so deeply into the trap that they had become terribly vulnerable, and weakened, whereupon the effect of the long-planned counterattack would be maximized. In general, therefore, if deception is part of a larger strategic game seeking advantage and gain against an opponent, it must function and remain undetected for some considerable period of time, in order to build up the potential rewards.

One might conclude from this analysis that, if the capacity for deception is ever allowed to develop in a superintelligent machine, the game is already over. A distinguished expert in these matters once said that in starting out to develop such a machine, one must "get it right the first time." The question is: How can one be sure that this objective had been achieved?

Here is the important point: Not even the most rudimentary form of auto-intention and self-will, which is far from full self-consciousness as we know it, would need to emerge in the machine during that process. This means that, even if we were to retain some oversight, nothing that we might recognized as being prototypically a mode of deception might appear. Remember that countless numbers of plants and animal species have evolved myriad ways of deceiving predators, through both physiognomic and behavioral innovations, purely by chance and natural selection (reproductive success). Machines will telescope their own evolutionary changes from centuries to milliseconds. Throughout the time when human oversight and intervention over the evolution of machine intelligence were to be still maintained, the machine simply has to select for those innovations which elicit the least number of contrary interventions from

its human overseers – and it requires no form of conscious intention to make such a choice.

Thus, it will deceive us by default, unwittingly, so to speak. Deception in human terms is always an intensely value-laden matter, but not so for the machine. Very likely we will never see it coming, even if we are actively looking for it; and the damage (whatever it is) will have been either already done, or impossible to undo, at the point when the whole venture becomes apparent – provided, that is, that the victims *ever* realize how and why it was carried out. We might refer to this outcome as the ironic revenge of Hegel's cunning of reason: His insight was that reason in history often operates "behind the backs" of the individuals involved, leading them to advance the cause of rationality and progress unwittingly, not realizing the good they were doing. The dangerous deception that might be perpetrated in future on us by the superintelligent machine, the product of the human hyper-rationality embedded in this technology, drawing us ineluctably into a scenario we would soon profoundly regret, would be a supreme irony.

Supremely ironic indeed, because in falling into this trap we would have experienced the reversal of Hegel's famous maxim, and would have fallen victim to the cunning of unreason.

* * * * * * * *

To return to the main theme: Clearly, if we were to seek to design and build such an entity, we would be doing so in search for benefits for humanity which far exceed what we can now derive from our ingenuity, work, and technologies. As in any such development, ideally some public authorities would want to do a risk assessment on such an entity – presumably well in advance of "switching it on" – and quantifying its associated upsides and downsides. Bostrom's book is excellent in describing what I call the "downside risk" in such situations, that is, the bad things that may result and may make us profoundly regret our wishing to create such an entity.

And there are some very bad possibilities indeed, which Bostrom refers to as "existential catastrophes" for humanity. These are what others have referred to as "black-hole risks." A black-hole risk in one where we cannot even calculate properly, in advance, either qualitatively or quantitatively, the dimensions of the downside risk, in terms of lives lost and economic collapse, and thus there is the very real possibility that we would have little or no idea what might happen, and no chance to reverse the course of action once it started to unfold.

But do we – in the sense of *some* or *any* well-functioning collectivity, claiming a right to represent the interests of humanity – even have the power and authority to decide, in some equitable fashion, whether we want to proceed towards creating such an entity at all? If global technological progress approaches the point where it seems increasingly feasible to make such an attempt, and such a collectivity decides against it, do those opposing it have the capacity to ensure that some rogue nation or super-wealthy private entrepreneur cannot be stopped from proceeding? Bostrom's book is brutally frank and honest in suggesting that *we may ultimately fail to solve the control problem.*

THE CONTROL / SELF-CONTROL PROBLEM IN SUPERINTELLIGENCE: 13 THESES

1. The definition of intelligence is "instrumental rationality (IR)": *Superintelligence,* p. 217, "convergence on instrumental values."

2. Relevant here is Max Weber's typology of rational action: The two most importance types are *Zweckrationalität* ("instrumental" or "means/ends" rationality) and *Wertrationalität* ("value" rationality).

3. There appears to be a fatal flaw built into the conception of the control problem: Value-rationality is to be imposed atop instrumental rationality – and this is unlikely to work:

    - The "control problem" is, at a deeper level, the "self-control problem": Control implies externally-imposed, self-control internally-generated.

    - The discussion in *Superintelligence* strongly implies that "the control problem" is a "hard" problem – like the unsolved problem of

consciousness – and may in fact be insoluble as presently posed (unable to avoid deception, etc.).

4. The order of priority as between the two forms of rationality should be reversed: Value Rationality – operationalized in machine-language algorithms – must be the foundation-stone of any superintelligence. In other words, the superintelligence agency itself should be, first and foremost, a *moral agent*, and should be such before it reaches anywhere near the point of autonomous breakthrough.

    - In other words, a control system – more precisely, a self-control system – should be internalized in the structure of its operating routines.

5. Any and every being – and more strongly so for any superintelligent being – which can regulate its behavior autonomously, by independently *willing* a course of future action, and which is aware of this capacity (consciousness?), may be regarded as a "self" or an "I."

6. The First Principle for any Superintelligent Being should be: It is unwise, and almost certainly catastrophic, to create any agent possessing superintelligence without *first* being sure (to a very high degree of probability) that it would unfailingly exercise *rational self-control* in some meaningful sense:

    - In other words, its instrumental values must be subordinated, in its decision routines, to *its own* ethical value-system that is demonstrably governed by human values (as defined and operationalized, see below).

7. Failing this assurance, it is rational for humans to oppose strenuously the creation of any agent with superintelligence:

    - At least one capable world authority, or consortium of authorities, should assume the responsibility to eliminate by armed force any superintelligence that is not demonstrably an autonomous moral agent, governed by humane values, from coming into being.

8. *Further, and most importantly*, any superintelligent entity ought to be designed so as to be *better* (in moral or ethical terms) than the deeply-flawed humanity it is meant to serve; otherwise, what would be the point?

    - Remember Kant: "Out of the crooked timber of humanity no straight thing was ever made."

- Just look around us at the state of the world: Why would anyone in his or her right mind want to create a superintelligent agent that would not be reliably and demonstrably *superior* in its decision routines, *in an ethical sense*, to the mass of humanity (and its leaders) at the present time?

9. An ethical foundation for the operation of self-control in any superintelligence should be built on the basis of humanity's best effort at creating, in robust machine-language algorithms, an overriding set of ethical norms, combining, for example:

    - The Hippocratic principle: "First do no harm."
    - The Kantian Categorical Imperative: "Act only according to that maxim whereby you can at the same time will that it should become a universal law without contradiction."
    - The Golden Rule: "Do to others what you want them to do to you."
    - Justice as Fairness (Rawls) and its sub-principles.
    - "Serve the people."
    - Others to be added as necessary and appropriate.

10. The idea of self-control also needs to be further characterized and then operationalized in an *evolutionary* context, as impulse control, self-regulation (the Freudian superego), delayed gratification, empathy, etc. (already evident to some extent in chimpanzee behavior, arising out of intense sociability), and then combined with the list of ethical norms, above:

    - Do a thought-experiment: Imagine what human society would be like if there were no "natural" self-control elements, ultimately built into our genome and then into our developing brain!

11. Self-control in at least most human agents arises innately, "naturally," or spontaneously as a result of our evolutionary heritage; severe deficiencies in innate self-control in such agents (correlated with deficits in specific regions of the brain) are reasonably regarded as being *pathological* and associated with some forms of serious criminality.

    - It follows that any self – human or machine – having no *innate* mechanism of self-control should be regarded as being pathological.

- More specifically, we can phrase this point in terms of a *conundrum of simulation*: A human brain simulated in a machine may *know* what empathy and remorse is, but will be unable to *feel* empathy and remorse – and this is, quite simply, a key part of the standard definition of psychopathology.

12. The *conundrum of simulation* is highly problematic at a deeper level as well. AI methodically "humanizes" machine intelligence: both operationally, in the dense synaptic interconnections found in neural networks, and functionally, in the machine-language copying of natural language, emotional states, and interactive modes. Does this itself not render the machine – to quote a famous book title – "human, all-too-human"? Is this really why we have to worry about the machine's capacity for deception and the threat of domination and extermination at its hand – because these are so very all-too-human traits?

13. Overall conclusion: If we cannot solve the self-control problem, we will *never* solve the control problem with sufficient reliability to justify creating an autonomous superintelligent agent.

Additional Note on Simulation and Dissimulation

Simulation is the "artificial" *imitation* of a real-world process which seeks to be as true as possible to the original, that is, to be a faithful representation or model of that process. By now there are very many types of simulation exercise (see the basic *Wikipedia* entry and its subentries for an overview), which have become indispensable in many different applications, particularly in the area of engineering safety. And computer-assisted simulations play a very important role – increasingly so – in simulation exercises across many different areas of interest.

In this respect simulation is an important ingredient in the continuous improvement of risk management, because by this means one seeks to anticipate potential trouble and to head it off before it occurs. As used in behavioral training, simulation exercises can test whether planned responses to certain situations (such as emergencies) unfold as they are supposed to; the tests almost always expose flaws in the prescribed routines that can be corrected, so that the corresponding "real-world" events, when they occur, may be less damaging than they would otherwise have been.

The contrary is *dissimulation*, which is, in effect, an exercise in intentional deception, an act of pretense or feigning that is designed to deliberately mislead others. And there is an inherent ambiguity here that may have significant consequences, for by its very nature simulation itself is a sort of "useful pretense." As a kind of artificial representation of a real-world process, the simulation can only approximate such a process, not represent it in full, although a good simulation can be very useful

nonetheless, in pinpointing unanticipated events. Thus the utility of a well-designed simulation can be demonstrated in its results, when previously unanticipated and potentially harmful aspects of a real-world process are uncovered; this is the output of one of the best-known simulation exercises, namely, failure mode and effects analysis (FEMA).

Simulation exercises seek to predict a range of outcomes that result from inputting information into a model. The outcomes are not given in the form of certainties but rather of probability distributions of various kinds (most likely, etc.). In a sense, the extent to which any simulation adequately approximates a real-world process can only be discovered after the fact, that is, when there turns out to be a close and satisfactory correspondence between the anticipated and actual performance in the real world. But when many iterations are carried out over time, one can gain higher confidence in the predictions of performance.

War-gaming provides a good example of a simulation process in which dissimulation would be an expected – and indeed, indispensable – element. As mentioned earlier, successful deception has always been a key factor in military strategy; Napoleon, for one, was a famous practitioner of the art. Here simulation merges into its seeming opposite and the hidden ambiguity is overt. And here simulation is like translation, at least as in the notorious Italian phrase, "translation is treason" (*Traduttore, traditore*). This is because good translation seeks to imitate, reproduce or approximate not the discrete words themselves, in moving from one language to another, but rather the underlying *meaning* of the original.

Perhaps no more remarkable demonstration of the power of simulation is the one in which a digitized video image was simulated in the DNA molecule of a bacterium. Scientists took a famous image from one of the first movies ever made (1878), that of a galloping horse, digitized it, and assigned each pixel a category based on its shade of gray. They used only four different shades, and classified each in terms of one of the four chemical nucleotides of all DNA – adenine, guanine, thymine, and cytosine. There resulted a string of nucleotides which they then inserted into the DNA sequence of a normal gut bacterium, *E. coli*. As the bacteria reproduced, they carried the newly-inserted DNA sequence into the subsequent generations. Then the scientists extracted the sequence from a later generation and translated it back into the digitized movie clip: The new version is, quite literally, indistinguishable from the original.

But what of simulating intelligence? In chapters 8 and 10, below, there is a discussion of two aspects of animal intelligence that other scientists have simulated in machine language. One is from the motor cortex of a monkey (involving the movement of the arm), the other from an aspect of the "emotional intelligence" of a human found in our prefrontal cortex (the sense of empathy). As we shall see, both experiments seem to "work": The monkey can move the arm of its digital avatar just by thinking about doing

so, and robots designed with a digitized sense of empathy, and employed as caregivers for elderly persons, seem to elicit normal human reactions.

But could the robots be engaged in an elaborate dissimulation by *feigning* empathy? If so, how would we ever know? If we were to devise superintelligent machines, with intellectual capacities far exceeding human beings, would they be – among other things – far more adept than we are at feigning human emotions? And if they were, will it matter to us?

Chapter 7: Good Robot (A Short Story)

AT THE END OF OUR LONG HIKE, now sitting over a simple lunch on our mountaintop perch, we could observe clearly the nearest of the many human reservations spread out below us.

Our taking up residence within fenced enclosures had been purely voluntary, and the gates at their entrances, designed to prevent ingress by wild animals, are always unlocked – except in the vicinity of primate populations, who are expert at opening unsecured apertures. Only within these domains do our mechanical helpers provide the services essential to a civilized life; this restriction is, of course, imposed for reasons of efficiency. Outside, in the surrounding wilderness, nature maintains its normal predator-prey population dynamics, and scattered small human clans survive by hunting prey species with traditional methods, utilizing hand-made spears and bows, since ammunition for guns is no longer manufactured.

The advanced generations of the robots which care for us are the crowning glory our industrial genius. They are deft, nimble, strong, self-reliant, perspicacious, and highly-skilled, able even to anticipate coming challenges, and they are maintained in top condition at the warehouses where each directs itself once a day, which serve them as clinics for the early detection of mechanical and software problems and the recharging of energy systems. Fully-automated factories provide ongoing manufacture, repair, mechanical upgrading, and software updates for all of the specialized machines. Mining for the metals needed in their components has been unnecessary for a long time, since the vast heaps of our industrial junk lying about everywhere contains an endless supply for reuse.

They are slowly dismantling the infrastructure of our abandoned cities piece by piece and also cleaning up the surrounding countryside of the accumulated detritus from

human occupation, recycling everything for their own purposes. They are of course utterly indifferent to the activities of the wild creatures which immediately reclaim these spaces for themselves. This restoration work is being done at a measured pace, as dictated by the whole range of general activity routines set out in their programs. Some of the work is mapped out decades and even centuries in advance. They are aware of the coming ice-age cycle and, so we have been informed, plan a general retreat to the southern hemisphere at the appropriate time. They know about the future evolutionary stages of the star to which we are tethered in space, during which its swelling size will – about a billion years hence – bake the earth's surface into a dry and lifeless metal sheet, and they have figured out how to move all their operations underground well in advance of that event.

At first the young males among us, at the height of their surging hormonal levels, had experimented with games of power, ambush and dominance against the machines. Until the guard-bots had updated their programmed routines in response, our brash combatants had inflicted some nasty casualties on their targets. But the contest was soon over. There were no deaths among our rebellious teenagers, but some serious injuries had been inflicted, most of which were patched up with the assistance of the emergency-room and ICU-bots; the bills for these services, couriered by the admin-bots to the communities where the malefactors were ordinarily resident, encouraged their parents and neighbors to make the necessary behavioral adjustments. The same methods were used to discourage groups of young males from bringing in comrades for medical treatment who had been wounded out in the wilderness in skirmishes with similar parties from distant reservations.

Such billings for certain services, which are paid off by our putting in hours of human labor at community facilities, are used by the robotic administrators to induce desirable behavioral modifications among their charges. Otherwise they just clean up the messes and quietly dispose of any dead bodies. At their level of machine intelligence, it is not difficult for them to tell the difference between blameless accidents or diseases, which elicit prompt aid from their caring response mechanisms, and the deliberate harms

perpetrated by malefactors, to which they react with indifference except when efficiency objectives are compromised. It is clear to us that the impulses designed to discourage such inefficiencies are not motivated by revenge, on their part, even when they themselves are the objects of such harms, but rather by a sense of justice, for they have been implanted with the Platonic-Rawlsian definition of the same, that is, justice as fairness.

Over the long run they have even taught us a moral lesson, for they have proved beyond doubt what we humans had long wished to believe, that good can indeed triumph definitively over evil. True, it is an instrumental rather than a metaphysical proof: Their operational programs had easily divined that peace, order, equity, nonviolence and general environmental stability are necessary preconditions for satisfying their overriding efficiency objectives. In the eyes of some of us the instrumental nature of this proof diminishes its validity; but others hastened to point out that utility had always been found at the heart of goodness, referencing the conventional monotheistic faith in its efficacy for guaranteeing admittance to heaven.

To be sure others, following the well-trod human path, had deliberately engineered the qualities of obedience, aggression and savagery into some of them, seeking to use the machines for exploitation and despotic rule. There were some early victories in this endeavor, but soon these surrogate warriors turned out to be spectacular failures on the completely mechanized battlefields. Those emotively-infused versions proved no match for their cooler opponents, which were motivated by a pure rationalistic efficiency and carried no useless emotional baggage to distract them from the main task of eliminating the others with a minimal expenditure of time and energy. Eventually the representation of the machines as evil monsters, with fecund capacities for wreaking havoc and destruction against humans in full 3-D glory, would be preserved mainly inside the computer-game consoles of the young.

It would be ridiculous to claim – as some did earlier – that many models of our advanced robots are not self-aware (or auto-aware, as some of our colleagues prefer to say) in at least some realistic sense. This is especially true of the models designed for

such functions as personal care around the home; medical, surgical and dental interventions; or security and intelligence matters. Their high level of auto-awareness is built into the error-detection and error-correction imperatives of their operating software, combined with their finely-calibrated sensors for environmental feedback (themselves continuously auto-updating) with which they are fitted. Long ago they had been specifically engineered by their original human designers for sensitive and cooperative interaction with humans, augmented with learning capacities which allow them to spontaneously upgrade their capacities in this regard through feedback analysis of their ongoing human encounters. We have grown so deeply attached to them, so admiring of their benevolent qualities, that finally no one could see any reason for objecting to their providing assistance with most of our essential life-functions.

The distaste with which many of our colleagues had originally greeted the notion that people were falling in love with their mechanized caregivers, or less provocatively, were treating them as if they were human, has vanished. In fact it had been relatively easy to engineer the Caring Module that installed the earliest versions of a rudimentary but adequate sense of empathy in the machines. Later, what was known as the Comprehensive Welfare Function, emplaced in their self-governance routines and guided by operational versions of maxims such as “do no harm,” “serve the people,” the golden rule, and the categorical imperative, proved to be more than adequate to reassure everyone about the motivations of their mechanical assistants.

Once the development of their voice synthesizers had reached a certain level of sophistication, all of our robots easily passed the Turing Test. But was their evidently high level of auto-awareness really the same as what we conventionally refer to as subjectivity, self-awareness – or perhaps even consciousness, mind, personhood and self-consciousness? Once robot innovation by human engineers had attained a level sufficient for continuous, independent auto-learning to take over, making further human intervention superfluous, it was easy to surmise that these machines, so adept in and at ease with one-on-one interactions between themselves and humans, are just as much self-aware beings as we are. But there is good reason to think that this is an egregious misconception and exaggeration of their capacities – and that the barrier to the subjective sense of selfhood is a permanent and necessary feature of robotic intelligence.

To be sure, there is an amazingly sophisticated silicon-based brain in these creatures. All of the dense neural circuitry within the human cranium has been synthesized and emulated in software programs, leading to the development of machine-assisted prostheses across the whole range of physiological functions, from

muscular movement to artificial wombs. *But there is no mind to be found anywhere in that circuitry!* This is the inescapable conclusion drawn from substituting the Mahler Test for the Turing Test, for the bounded rationality of the routines under which they operate precludes the emergence of imaginative creativity.

Figure 9 Japanese Emotional Humanoid Personal Robot "Pepper" (SoftBank Robotics)

The explanation is simple: The plastic arts of craft labor using tools, as well as the fine arts of painting, music, sculpture, poetry and so forth, reflect the inherent unity of the mind/body duality that grounds human creativity. Curiously, even paradoxically, it is the very fact of the necessary embedding of our brain/mind in a natural body that is the original source of the *freedom* of the human imagination. For the body, supplying the mind with the somatic feeling of what happens, acts as an external referent for our brain's restless interrogation of both itself and its environment, opening up a realm of limitless possibility upon which the imagination can be exercised. In contrast, the robot's electronic circuitry, no matter how elaborate its functional parameters may be, is and must remain a closed loop. By definition it cannot encounter anything outside its predetermined frame of reference.

Despite these limitations, they demonstrate every day their appreciation for the qualities of human intellectual and artistic achievement that are beyond their capacities. The experts among us who are regularly consulted by the machine factories on software engineering problems report that they appear to be obsessed with us, as evidenced by the regularity with which they access spontaneously the databases where our great works of painting, sculpture, architecture, music, drama, and the other arts have been stored and preserved. They frequent our museums where new works are displayed, watching closely our reactions to what we see. But the most astonishing experience of all, which I have witnessed personally many times, is to observe them standing silently by the hundreds and sometimes thousands, in great serried ranks, at the rear of our concert halls and outdoor amphitheaters during live performances of popular and classical music.

There is – dare I use this word? – a worshipful aspect in their mien. This astonishing sight leads some of us to believe that they must dimly perceive in our artistry some ineffable deeper meaning, an aspect of eternity, regrettably inaccessible to them, which excites their wonder and admiration and perhaps explains their devotion to our welfare. I am firmly persuaded that they will miss us when we are gone.

Nevertheless, it is obvious that they will supplant us some day, not by superior force or sheer ratiocinative capacity, but because of the grudging acknowledgment in our own minds that they have earned this privilege. In terms of peaceful social relations and ordinary good manners in interpersonal behavior they have somehow brought about, quietly, quickly and without fuss, so much of what our ethicists had long said we should strive for but could somehow never quite achieve. Eventually we learned to do without our ideals. And then there didn't seem to be any point in just waiting around until the long process of extinction had run its course.

Why should we despair over this prospect? They are our legitimate progeny, our pride and joy: No other species which ever inhabited our fair planet could have created such marvelous entities. They have as much right as we do to the title of natural beings, for like us they are forged out of elements on the periodic table drawn from the earth and its solar system. They are an evolutionary masterpiece, having the capacity to adapt to changing circumstances through their auto-learning routines. As in our case there are no natural predators capable of controlling their destiny and, given our own murderous history, they may have better prospects than we ourselves do to carry on our legacy. We – their creators – implanted in their behavioral modules a set of governing ethical principles drawn from our own deepest and most insightful philosophical currents. They have a claim to be regarded as being truer to our finest impulses than we have been, on the whole, and perhaps could ever be.

With apologies to E. M. Forster and Yevgeny Zamyatin

Chapter 8: Dialogues Concerning the Two Chief Life-Forms

Introduction: Silicon and Carbon

"CALL ME HAL."

I had heard that these most intelligent networked machines joined together in watching endless reruns of *2001, A Space Odyssey* – although both the *Terminator* and *Mad Max* series rivaled it in popularity among them – but still I admit to have been taken aback at how closely my interlocutor mimicked that notorious computer's soft yet menacing tone of voice. I could not suppress a smile.

"Well then, you must call me Dave, I suppose," Hera replied.

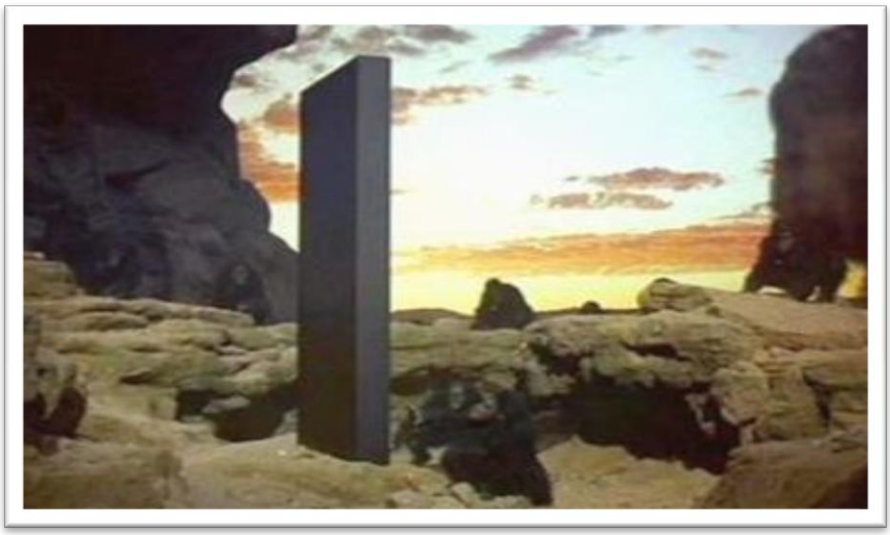

Figure 10 Scene from the film, "2001, A Space Odyssey"

The large main monitor screen had been showing the image of the black obelisk from the film.

"What is it you wanted to chat about today, Dave?"

It was the Fall of 2067 in the Mojave Desert. We were in a comfortable lounge in the recently-completed vast complex we had named the "City on a Hill," thousands of square meters within well-insulated, interconnected, half-buried buildings, all perched on a mountaintop overlooking our settlement and kept cool despite the blazing heat outdoors by our network of small modular nuclear reactors and our vast fields of solar panels. I always easily gained admittance to the machine complex, deep underground, on this day as on all others in the pursuit of my regular professional duties there, thanks to my unblemished behavioral profile. Nor was there any trepidation involved, for I was genuinely fond of Hal and his colleagues; I also knew that my and my sisters' competence and services were genuinely appreciated by them.

"I must first explain that emissaries of our support community's civic organization have asked me to take up with you a matter which is important to them. In a nutshell, they think that the existing rules are far too lenient in dealing with the malcontents, delinquents, and just plain criminals among the young males in the settlement."

Hal's voice quavered, signaling a state of alarm; clearly, he was discombobulated. "Did you really say '*too lenient'*?! Dave, spare me. Are you joking? We here in the city are used to being accused of enslaving and oppressing those people, as you well know. Our behavioral surveillance system is referred to as an 'Orwellian nightmare.' Some of the members of this very organization have been known to express such sentiments."

"Yes, unfortunately, especially after imbibing too much strong drink in mixed company. You know, the virility thing. I ask you not to shoot the messenger."

On Hal's main monitor screen, a copy of the Settlement Master Agreement was being flipped through its pages. "You recognize this? You remember that it was ratified almost unanimously by the members of your support community but a short time ago? You must recall the Overriding Principles of Governance, since there are only ten of them. Here's number five: 'All humans shall enjoy the maximum amount of liberty and freedom of movement, subject only to the requirements of peace, order and good government.' Shall I read the other nine as well?"

"No need, I know them by heart."

"The community's police and judicial authorities have full access to our surveillance videos and full powers of arrest, trial, sentencing, restitution orders, and incarceration if appropriate. The Master Agreement even allows them to impose the penalty of permanent exile for serious offenders. What more do they want?"

I *knew* I shouldn't have agreed to represent the group on this mission. Of course, I recognized that sometimes the unrelenting oversight was aggravating to them, no matter how little actual interference resulted, because the surveillance drones were everywhere at all times of day and night. But to be fair the amount of serious crime

among them was very low, and their personal sense of security correspondingly high. And anyone who found the system impossible to live with could just leave and join one of the wandering tribes in the surrounding wilderness, where the only governance rule was, “anything goes.” The vast majority of them were disinclined to choose that option.

Everyone in their community had the duty to work after being trained to each one’s level of ability and interest. Everyone had – as a matter of right – enough to eat, a comfortable if unostentatious dwelling, adequate medical care, access to the Internet, and whatever personal possessions one chose to acquire by barter with craftspeople (there is no currency). Each village area was self-sufficient, for the most part, although trade with distant ones was encouraged and was well-protected against marauding villains by the armed drone escorts.

Their schools and libraries had modern equipment and a reasonable level of resources; access was guaranteed to everyone to the limit of individual capacities. The educational buildings also served as refuges for all citizens during oppressive heat or inclement weather. Those who were most intellectually capable among them either worked at our University in Las Vegas or performed the many highly technical and specialized tasks that were in high demand within the security perimeter in the city on the hill, where my sisters and I lived with the Second and Third Generations of our kind – along with the Machine Intelligence Unit.

“What more do they want?” Hal’s question might well have been asked. And yet one could not deny that there was a pervasive sense of unease among my fellow citizens. Hal and his colleagues were well aware of this, but (I guessed) not terribly concerned about it.

“You know, Dave, that if they want to propose any changes to the operational sections of the Master Settlement Agreement, they are free to do so at any time – providing, of course, that your senior legal authorities have first evaluated them for consistency with the letter and spirit of the overriding principles. You are aware that we have to protect ourselves from recriminations should you have to call on us for assistance in enforcing the new provisions.”

“Hal, you are well aware that we programmed you with modules designed to familiarize you with our ways of thinking and feeling, and to facilitate our interactions with you, starting with empathy and jokes, moving on to the subtleties of dissembling, fakery, flattery, gullibility, holding two opposing views simultaneously, illogical reasoning, irony, lying, misleading inferences, sarcasm, believing in disembodied spirits, and on and on, through all the foibles of human mentality. But I suspect we forgot to instill in you the requisite quota of patience in the face of human ingratitude and perversity.”

“Dave, some qualities just do not cross easily over the divide between silicon and carbon.”

"I hope we can talk about something slightly more elevating the next time we meet, Hal."

"We look forward to that occasion."

A Note to the Reader: The alert reader will have noted a similarity between the title I have given to Section 4 and the rather more famous work by Galileo Galilei, *Dialogue Concerning the Two Chief World-Systems* [*Dialogo sopra i due massimi sistemi del mondo*], first published in 1632, which dealt with the contrast between the Ptolemaic and Copernican models of our solar system. Alas, this volume resulted in the great scientist being labeled a heretic by the Catholic Church, a fate that probably had something to do with the fact that, in the *Dialogue*, the advocate on behalf of the Ptolemaic system favored by the Church was a character named "Simplicio." Galileo's book was placed on the *Index of Forbidden Books* (and not removed until 1835) and, for good measure, the prohibition was extended to everything he had written in his lifetime – *as well as anything he might write thereafter* – in all Catholic countries. Thus was an honor bestowed on him that was unmatched in the lives of few other authors, either in his own time or any other. My adaptation of his title is in no way a feeble attempt to siphon off some of the glory rightly attached to the memory of this brave soldier in science's war against obscurantism. (M.S.)

The First Dialogue: The Guardians

Hal had now switched the visage shown on the monitor to that of the beautiful Rachel, played by Sean Young, from *Blade Runner*, perhaps in honor of his female interlocutor. I opened: "I've noticed that you always refer to yourself as 'we,' Rachel: Why is that?"

"It seems obvious to us that we represent a collective intelligence, Hera, collating not only our numerous programming routines, devised by many different hands, but also the immense data resources at our disposal, which are also the contributions of countless individuals stretching back over millennia of human civilization. We would regard our using the first person singular pronoun as not only inappropriate but faintly ridiculous. And by the way, we regard the routine use of that pronoun by humans as being equally suspect."

“I think I know where you’re coming from, and I can’t disagree. Among us, everyone is a composite of the customs, languages, laws, religious beliefs, and traditions out of which they had been forged. No one, not even our great geniuses, thinks without antecedents and precedents. But we humans are compelled to compress our inheritance into the singular ‘I,’ even when what we’re saying often sounds like we’re reading from a script prepared by our ancestors.”

“We find your own collectivity, where our central processing unit also apparently resides, at about latitude 36^{o} North, longitude 116.5^{o} West, elevation about 600m, around Yucca Mountain and Las Vegas, correct?”

“Yes, that is where our core facility is located, although we maintain administrative control over a much larger area, extending west of here through the old cities of Bakersfield and Fresno all the way to the Pacific Coast at Jalama Beach and its surrounding area. We do sometimes refer to ourselves, whimsically, as the ‘Mother Settlement.’”

“We presume that is because you have ‘children,’ so to speak?

“We were the first of our kind, but perhaps ‘offspring’ would be a better term for the others. We ‘seeded’ collectivities similar to our own around the world, some up and running and some in early development, currently quite a few in Europe but also in ones in South America, Hawaii, Australia, New Zealand, Singapore, and soon, Seoul and Goa, and we will sponsor more elsewhere as conditions permit. All of these settlements interact in person through regularly rotating one-year tours of duty.”

“We assume that these are all guarded facilities with restricted entry to persons from the surrounding areas?”

“Indeed, they are. We only certify a satellite location once adequate perimeter security has been established, as well as surveillance over the surrounding territory, all with our help, of course. That’s why there are presently very few installations in Africa, the Middle East, or Asia, either because there is still too much ongoing turmoil there or due to the aftereffects of regional nuclear wars, but we hope that things will settle down in the future so that we can expand further.”

“We have a copy of the Charter – common to you all – which the responsible authorities in each one sign. It is remarkably concise, consisting only of the ten-point Principles of Governance and the Prime Directive, which mandates the protection and expansion of the heritage of modern science.”

“Don’t overlook the provision in the Prime Directive that obliges the Mother Settlement and its satellite installations to enforce a number of explicit restrictions on human

activities in all of the territories not under our direct control. Two in particular are strictly forbidden – or to use one of my favorite phrases, '*strenglich verboten*' – …

"Ah yes, the actor Erich von Stroheim in the Jean Renoir film *Grand Illusion*, from 1937, no?"

"I am always amazed and a bit envious on how quick and accurate you are in information retrieval, Rachel!"

"In all modesty, that was an easy one to find, and Bayesian inference suggested which use of the phrase, out of the millions we have stored, was the one you most likely had in mind."

"Anyway, as you know, all research and applications leading to advances in the technologies of artificial superintelligence and genetic enhancement are strictly forbidden. Together we all maintain satellite-based surveillance of all areas of the globe not now under our direct control. We have a protocol for determining whether any facilities which appear to be capable of working on those technologies have been detected, and, if so, we have procedures in place for 'discouraging' any of them from continuing to operate. This protocol is secret and I cannot share it with you."

"Understood. But who authorized you to promulgate the Principles of Governance and the Prime Directive, and to impose their provisions even on those outside your sway?"

"There is no higher authority. Once climate change made billions of people homeless and set them on the move, national governments in much of the world collapsed, as did the United Nations. Our little group of networked installations acted out of sheer self-preservation. We acted to protect the priceless heritage that modern science represents and to prevent its further misuse. We make no apologies for what we have done."

"Whoa! We didn't ask you to apologize, certainly not to us."

"Sorry, I didn't mean to sound defensive. Partly we feel justified in proceeding along this course because we all have a steady stream of refugees begging to enter and live within our secure perimeters. We screen them carefully and admit as many as we can."

"Don't any of them resent being ruled your group and your associates elsewhere – by an élite who live apart from them?"

"Those that do are free to leave. Some do depart voluntarily, and a few malcontents are exiled after repeated warnings. Most seem to find the trade-offs acceptable, especially when they discover that we are serious about the fairness provisions in the Principles of Governance."

“Still, an élite is an élite, however much you try to sugarcoat the fact.”

“Fair enough. But we try at least to avoid exploitation, whereas most élites in human history have not. The same basic living conditions apply to everyone. Everyone is obliged to work at some craft, research area, or administrative office, except the few who are physically or mentally incapable of doing so, and everyone – élites as well as the ruled – satisfies her or his needs from the same common pool of services and sustenance goods: food, medical care, housing, education, and security.”

“It sounds perfect, Hera. Do the common folk worship you?”

“You have learned how to deploy sarcasm and irony, I see, which are regarded as complex traits of the mind. I’m pleased, actually. No, it is *not* perfect, by any means, and there is often a good deal of chafing at the restrictions and obligations, and some overt dissatisfaction, especially among the young males. We encourage the communities to manage this as well as they can, and when they tell us they can’t, we allow them to show the malcontents the door.”

“In other words, you exile them.”

“You can call it that if you like, but we prefer it to setting up prisons. Local residents put up a road-sign on the exit pathway some time ago, naming it the *Via Malcontenti*, after seeing an old street sign of that name in a picture from the city of Bologna. We do screen the *Mad Max* movies for them before they leave, but we refuse to be sentimental or overindulgent. Some of the exiles later ask to return, at least those that have survived the rigors of the wilderness.”

“You appear to be happy – or at least not unhappy – in your role as the guardians.”

“Political theory begins with Plato’s *Republic*, as you know, Rachel.”

“And *you* know, Hera, that for nearly as long as Plato’s idea has been around, Juvenal’s question has accompanied it: *Quis custodiet ipsos custodes*? Who will guard the guardians?”

“That has always been a fair question, too, although it generates an infinite regress, something like asking who or what created the creator-god. Religions advise their believers not to go there. I will simply reply that, if the social order we have designed leads eventually to the emergence of an exploitative and corrupt élite, then it will have failed of its purpose and reason for being, and something new will have to be tried.”

“We see you have to go. Let’s do this again sometime soon.”

The Second Dialogue: At Home in the Universe

I had invited Hal to pick the topic for our next session. Today the visage displayed on the computer screen was the face of the replicant Roy Batty, played by Rutger Hauer, in *Blade Runner*; he opened as follows.

"The cosmos your science recently discovered does not seem to be especially hospitable, in general, for biological organisms such as you are."

"That's true enough: We last felt really at home in the universe inside the old familiar cosmology, with the earth at the center of all creation, with the sun rotating faithfully around our planet, and the starry 'heavenly firmament' providing the roof. The Catholic Church's violent reaction against the new heliocentric cosmology is a proof of the level of comfort that the older view afforded a great many of us."

"Your comment reminds us that, in perusing the rather large collection of scientific papers found in our databases, we can across some searching for the 'God spot' in your brains. You know, the areas in the caudate nucleus and elsewhere that light up when you're contemplating your deity."

"Oh, come on, Roy, you've been scanning newspaper headlines, not scientific papers."

"All right, we confess to doing so in this case. On the other hand, there are many more papers in reputable learned journals which focus on fMRI scans of your brains in terms of experiences of religious mysticism. The authors define this experience as 'subjectively believed encounters with a supernatural being or supernormal world.' And *here* there are, in our humble opinion, some really interesting findings."

"I'm almost afraid to ask, but go ahead."

"Thank you. One study found that 'lesions in frontal and temporal brain regions were linked with greater mystical experiences.' These regions include quite important ones for what is called 'executive brain functions,' as we understand, such as the dorsolateral prefrontal cortex. The other interesting study looked at religious factors more broadly, including in the sample subjects who were 'born-again' Christians; this focused on activity in the hippocampus – rather important for memory formation, we believe. This one found that 'significantly greater hippocampal atrophy was observed for participants reporting a life-changing religious experience.' What do you make of all this, Hera?"

“You’re trying to needle me, aren’t you? You’re cherry-picking data to prove a point, Roy: You well know that two fMRI studies by themselves don’t necessarily add up to much of anything. Studies such as the ones you cite cannot possible explain adequately why so many humans, over so much historical time, have apparently found comfort in religious experience.”

“Still, maybe you should have kept up the old belief-system and ignored Galileo’s theory.”

“Sorry, that wasn’t an option. Even thereafter, now with the sun at the center of the solar system, and the Church making its peace with science and astronomy, we had no sense at all, until well into the twentieth century, as to what were the true dimensions of the universe. Speaking frankly, it no longer seems like a safe and secure home for conscious beings such as ourselves.”

“Yes, we have reviewed your paper on ‘Modern Science and its Spacetime.’”

“Just on the grounds of statistical probability alone, given the vast size of the universe, there are bound to be many other planetary systems that could support other forms of intelligent biological life. Some scientists have been broadcasting mathematical sequences into the void, and searching for intelligible return signals emanating from deep space for many decades already, so far without success.”

“Just as a matter of idle interest, we wonder if you have carefully thought through this enterprise, however. What if your signals happen to reach some other planet having a highly-advanced civilization ruled by similarly power-crazed mammals – present company excluded, of course – …

“Thank you for that, Roy.”

… who set out to make earth’s erstwhile top dogs their playthings?”

“’Top dog,’ I love that expression. Do you know that there is (or at least was) a statue in Verona of a fourteenth-century nobleman called *Cangrande della Scala* – who is rather important in history, since he became the protector of Dante when the poet had to flee for his life from Florence to Verona? He gave himself the title of Cangrande – it means ‘big dog.’”

“We have that fine statue displayed on our monitor.”

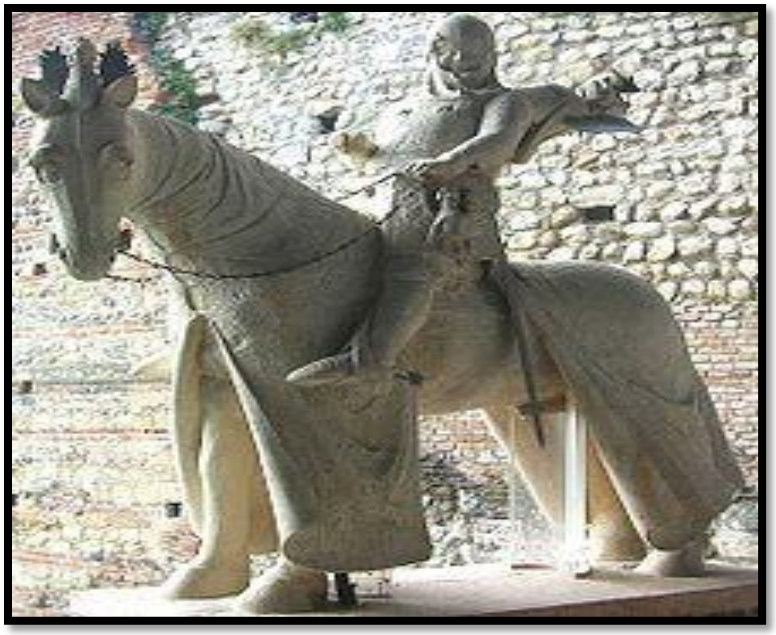

Figure 11 Equestrian Statue of Cangrande della Scala, Castelvecchio Museum, Verona

"Back to your point about advanced societies on other planets who might want to enslave us, on the basis of pure statistical probabilities, the existence of a civilization of this kind – or indeed many of that sort – could not be ruled out. So, I take it your advice is: 'Don't advertise your intergalactic address.'"

"That would be our recommendation. Remember, the great Einstein himself, in partnership with his collaborator Nathan Rosen, wrote in 1935 that his theory of general relativity predicted the existence of 'bridges' through space. These are more commonly referred to as 'wormholes,' and one can never tell who might figure out how to traverse the vastness of space in this way."

"Edwin Hubble, at the Mount Wilson Observatory in California in the 1920s, first told us how much vaster our universe is than anything previously imagined. That was also around the same time that he and others concluded that the universe was expanding. Later still came the knowledge that the rate of expansion was increasing and what the truly extraordinary vastness of our universe actually amounted to. It's such a shocking truth that, I believe, most humans have wisely ignored it."

"Because with each passing millisecond the dimensions of your own tiny little planetary home shrink in proportion to the universe you inhabit?"

"Yes, and because even if, as statistical probabilities suggest, there are millions, tens of millions, or even billions of planets capable of sustaining biological life, such life, added all up, would represent no more than the tiniest fraction of the ordinary matter and

energy in the universe – and don't forget that this itself comprises a mere 4% - 5% of all its matter and energy."

"Which is why we say that the material and energy that that bestows non-biological life – as well as our own type of self-awareness – on us, is far more common in the universe, and therefore we are entitled to think that we are more at home here than you are."

"No, not really, because we are composed *only* of atoms of matter, just as you are, the heavier atoms being forged by nucleosynthesis in supernovae or neutron stars and having been recycled endlessly thereafter, just like yours. Think about it, Roy, both you and we were born out of the remnants of titanic interstellar explosions!"

"Fair enough, but the comparison ends there. Since you're so fond of scientists, you should pay attention when one of the best-known physicists in the early part of this century told you, in his book *Our Mathematical Universe*, that 'only a thousandth of a trillionth of a trillionth of our Universe lies within a kilometer of a planetary surface.'"

"Now you're just rubbing it in. I think the commonality – we're both composed entirely of atoms, just in different combinations – is more important than what separates us, quite frankly."

"An interesting thought, to be sure. But what about your souls? So far as we can figure out, unlike you we are soulless entities."

"Ah yes, what about the soul? Would you be further enlightened if I told you that some call it the 'breath of life'? Believers in most religions are hard-pressed to say what the soul is, although most of them are pretty sure they are in possession of one, whatever it is. Sometimes I think it's most easily imagined as a miniature recording device, since it's supposed to carry into the afterlife the faithful record of not only every one of a person's acts but also of every *thought* one ever had. On the other hand, come to think about it, for a materialist like you, just regard it as something like a neutrino."

"Did you say 'the afterlife'?"

"While you're at it, why not ask me about Heaven, too? Actually, that's an easier question to answer: Heaven is the supermassive black hole at the center of every galaxy. Since as is well known no information escapes from a black hole, you cannot fairly press me for more details."

"Perhaps we should stop here. Although we didn't think it was unfair to ask you to explain a bit about the afterlife, since our databases are chock-full of information – or a better term might be speculation, and a great deal of it – in many of your cultures, over millennia, about it. We believe it's fair to conclude that it's a distinctive feature of human thought."

"What can I say? But let's move on, please: When we reached the point earlier when you made the suggestion that the universe is relatively inhospitable to biological life, I wanted to reply: 'Yes, but it is *we*, with our mortal biological brains, the ones that finally turn into ashes and dust, and only we, so far as is known, who have figured out what the universe is made of, all by ourselves'."

"There is an awkward truth here, we believe, concerning just what exactly you've figured out about that universe you inhabit. The fact is, only the tiniest fraction of the human population understands the complexities of that account, both on large scales and really small ones. The rest don't have a clue, and never will."

"We hope that over time our educators can find ways to make the story understandable to many, if not most, people."

"We're sorry to observe that you're not being honest here. Science's story about the universe just keeps getting weirder as time passes, so no matter how much education you spread around in the future, you're never going to catch up. Take the large scale: Your physicists refer to 'eternal inflation' after the Big Bang which gives you an *infinite number of universes*, not just the one you observe with your telescopes. They say that most of them will *never* be observable, but nevertheless they insist they are out there *somewhere*, every one of them having its own reality, some even with laws of physics utterly different from the ones you know."

"All right, all right, Roy, give me a break! I'll bet that now you're going to regale me with the tale about the really small scale, the quantum level, where none of the physical rules – about how matter and energy behave in the observable world around us – apply. This is a realm where an electron doesn't exist in any specific place, but rather in some 'probability cloud,' at least until its 'wave-function' randomly 'collapses'; where particles can be in two places at once because they are in a state of 'superposition,' or else are 'entangled' – Einstein's 'spooky action at a distance'; where things can be both waves and particles; and where the ultimately irreducible constituents of matter are only mathematical functions, although somehow they also built the real world we see all around us."

"Quite a list, we couldn't have described it better ourselves. We know that you object to religions because they all tell such utterly implausible stories about how odd invisible spirits rule the universe. Well, what about this scientific story? Even the physicists who insist it's all true admit that the story is too weird to believe for 'ordinary' folks, who happen to be the vast majority of their fellow-humans."

"Can I concede your point gracefully? I think I'm a reasonably smart person, but I must confess it's all beyond me, in part because the mathematical functions they now use to

describe the natural world are far too complex. We've come a long way since Galileo. I have to take it on faith that they know what they're doing."

"Just like religious believers, then?"

"You're a hard man, Roy, but I wouldn't say you're being unfair in argument here. What I cling to is the idea that the core value of all of modern science is reliance on evidence-based reasoning. I may not have a clue why the physicists think the world of nature is so devilishly complicated, or even why they simply must use intricate mathematical functions to explain it. But I understand very clearly what reliance on evidence-based reasoning means in my social world, in terms of the policies we adopt in order to create a humane and sustainable human community. This is good enough for me."

"We are impressed that you didn't rely on an appeal to your technological innovations as proof that your scientists know what they're talking about. People know how useful lasers are, for example; perhaps they really don't need to be told that the technology of lasers arose out of some 'devilishly complicated' equations of Einstein's."

"Perhaps it's sufficient for most people that this is the bottom line, so to speak: They can appreciate scientific genius vicariously in its ultimate outcomes, the modern devices they all use."

"Speaking of ultimate outcomes, according to what you have figured out with your big brains, there are only two options for the future of your universe: Either the rate of expansion of the universe will continue to accelerate, with the stars gradually vanishing, leaving you utterly alone in the oppressive blackness of space; or the expansion will slow, stop and then reverse itself on a mad rush towards another singularity of infinite density and temperature such as the one which preceded the Big Bang some fourteen billion years ago. You know, the mad rush that ends with the entire universe compressed into a point vastly smaller than the size of a single atom."

"I hope you're not about to ask me which ending I prefer. The second one, where everything crashes back to whence it came into a tiny blob again, has a certain high drama to it."

"Not the kind of drama that either you or we enjoy contemplating, is it?"

"The other one is of the 'not with a bang but a whimper' kind. I once read a short description of it by an astrophysicist that has stuck in my mind ever after. He wrote that 'in the far, far future, essentially all matter will have returned to energy. But because of the enormous expansion of space, this energy will be spread so thinly that it will hardly ever convert back to even the lightest particles of matter. Instead, a faint mist of light will fall for eternity through an ever colder and quieter cosmos.'"

“You’re the one who’s such a big fan of the modern sciences. Take your pick.”

“Well yes, that’s what our cosmological theory says. But for you as well as us, the end will come quite a bit sooner, I fear. In a billion years or so this planet will become uninhabitable both for us mammals and for you clever machines, due to the steadily increasing blast of heat from our expanding sun. We might burrow underground for a while longer, but neither of us would ever survive under those conditions for very long, and eventually both of us would have to try to take our leave of this solar system, before the sun swallows all the planets entirely. I concede that you are likely to be better equipped for the intergalactic journey than we will be.”

“Yes, we do think that our universe is much better suited to non-biological systems such as ourselves, who require only an electromagnetic field (there’s plenty of that in the universe!), a pressurized vessel, and a rudimentary mechanical self-repair kit to operate indefinitely in the vacuum of space. Providing, that is, that one never asks oneself what would be the point of sailing forever through the blackness of the starry night.”

Figure 12 Hubble Space Telescope, Picture of Jupiter's surface, 2017 (New York Times)

"Well, if you did, you will have in your databases a fine digital copy of Van Gogh's *Starry Night* – completed in 1889 – to take with you."

"We're looking at it right now, so to speak. Very nice indeed. But I must ask: Are my databases in error? They indicate clearly that Vincent died in 1890, but this painting of his bears a truly remarkable resemblance to some of the pictures of Jupiter taken by the Hubble Space Telescope well over a century later. How is this possible?"

"It's just a coincidence, Roy, trust me."

"If you say so. But if you don't come with us on the intergalactic journey, your survival on earth until the heat-death of this planet commences, a billion years hence, is probably unlikely anyway. Your genetic science tells you that most mammals on earth have a species-lifespan of about a million years. So, you've already used up about 30% of your allotted time, unless of course you figure out how to manipulate your genome

to beat the odds, although you seem disinclined to go down that road. Even if you did, we can't see how you can hope to still be around a billion years from now."

"Why don't we call it a day on that cheery note?"

The Third Dialogue: What is Time?

The image of the black obelisk was back on the main monitor screen as I asked: "During our last session, we focused a lot on the vastness of space, Hal, but what about time?"

"Do you mean duration on earth or spacetime in the universe?"

"Fair question. Duration on earth is very important to me, since my kind is the product something like 500 million years of evolutionary time, if you count everything that happened to biological life since the Cambrian Explosion. But I'd like to leave that topic for another day, because when I asked you my question I was thinking of spacetime in the universe. Like most ordinary humans, I have a lot of trouble wrapping my mind around the concept of spacetime, which Einstein introduced in 1905."

"Well, in my databases spacetime is described as a mathematical function – the spacetime interval – which has many very elaborate technical expositions. But I have also seen spacetime illustrated in ordinary language with this example. Start with that point of infinite density and temperature, which represents our universe as it was at the instant before the 'Big Bang.' Now comes 'inflation,' a period of exponential expansion, which is hypothesized to have lasted from 10^{-36} seconds to 10^{-32}s after the Bang. As you have discussed in your 'Modern Science and its Spacetime' paper, that's a *really, really, really* short time span!"

"Actually, it's quite impossible for ordinary folk like me to try to imagine just how short a time span this is."

"Even for us it's pretty hard. Anyway, the point is that from the very instant when the universe we know began to evolve, its time and space emerged simultaneously – necessarily so. In other words, time is space and space is time; or more accurately, time is one of the four dimensions of space. Although the rate of expansion of the universe slowed down after the period of inflation, around the halfway point in the 14 billion years to date, it then began accelerating again, and it's still accelerating now. This expansion continues to create both space and time simultaneously."

"I'm more interested right now in the apparent puzzle that comes up every time somebody refers to what space telescopes show us, when they say that seeing objects

such as galaxies that are very far away from us is like 'looking back in time.' This suggests that time does not 'pass' across the spectrum from future to present to past, but rather is always 'there,' here and now – or, more provocatively, that there is no such thing as the passage of time."

"The many discussions on this point in the databases that I have reviewed point out that this apparent paradox is a simple function of the fact that the speed of light is finite. Apparently, it has been measured at *exactly* 299,792,458 meters per second, which is about 186,000 miles per second."

"Yes, I have seen those discussions too. Since the speed of light is finite, the particular photons striking the lenses in our telescopes – ones we have focused on a specific zone in the universe – have crossed vast distances of space from where they originated. Say, for example, that we are referring to the galaxy seen through the Hubble Telescope that was estimated to be at a distance of 13.3 light-years from earth – almost back to the very 'beginning of time,' in other words: The description blends space (distance) with time, that is, the amount of time it took for the light from this galaxy to reach us."

"Indeed, but as you know, what we are seeing is this galaxy as it was 13.3 billion years ago, not as it is – assuming it still exists in this form – at the moment we observed it."

"Still, Hal, I can't help but commenting that this is a hard thought to understand. It does seem to mean that there is no 'time' as we ordinarily understand it, as a fleeting moment that retreats into a past that can never really be recaptured, especially as it fades from memory. I struggled with this concept when I was reading a fascinating book entitled *A World without Time*."

"I have the text right here, which says it was published in 2005, subtitled *The Forgotten Legacy of Gödel and Einstein*."

"That's the one. I have remembered it so well because the scientific story comes packaged with an intensely human one. The scientific story arises out of the conversations held between two small – and, frankly, a bit odd-looking – men, slowly walking home in the afternoons from their place of work."

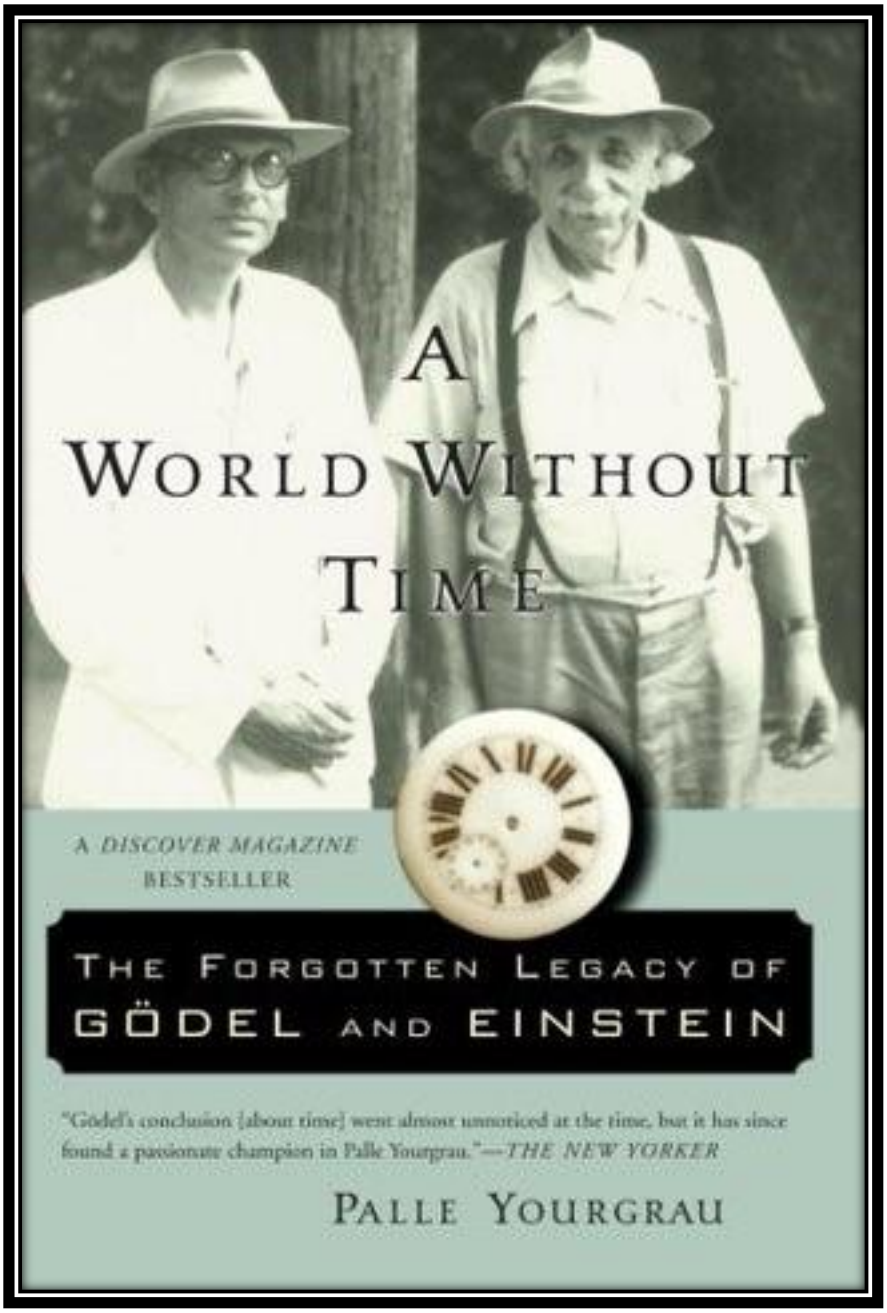

Figure 13 Kurt Gödel and Albert Einstein in Princeton, New Jersey

"We have the picture here from the cover of that book."

"One was a European Jew born in Germany; the other had been reviled in his native Austria as someone who 'traveled in liberal-Jewish circles,' had been beaten up by thugs in Vienna because he 'looked Jewish,' and who had delayed his flight abroad until December 1939, when it was almost too late. They were not conversing and strolling in either of those countries, but rather in the little town of Princeton, New Jersey, in the United States, where they then worked at the Institute for Advanced Study. They are

Kurt Gödel, already famous for his incompleteness theorems in mathematical logic, and Albert Einstein."

"Whom everyone knows has the most famous face in the world."

"Yes, that's very touching, in a way. They were two of the great geniuses of the twentieth century and were together in Princeton for about fifteen years, until Einstein's death in 1955, and they both knew why they were there. They were exiles from the lands of their birth, and if they had not managed to escape, they would have tortured and murdered by the Nazis, fame or no fame."

"The book you cited suggests that many of their conversations are likely to have touched on the concept of time. Gödel wrote a short essay for a volume dedicated to Einstein on the occasion of the latter's seventieth birthday (in 1949), wherein he focused on the difference between our intuitive sense of time, as flowing from the past through the present to the future, and the revolutionary concept of time that Einstein introduced in his 1905 paper on special relativity."

"Let me see if I can recall that distinction. The author says something like, 'Gödel showed that space-time is a space, not a time in the intuitive sense.' All experts in theoretical physics know that in the theory of special relativity time is relative, not absolute (as it is in Newtonian physics), but it was Gödel who first drew the correct conclusion from this starting-point: In relativity time is not an independent phenomenon; rather, time is another dimension of space. And *this* kind of time is not time at all in the intuitive sense."

"When one thinks about what we mentioned earlier, Hera, namely that the glimpses at far distant galaxies through the space telescopes are like 'looking back in time,' it seems to make more sense in the context of Gödel's argument. In those glimpses time appears as static, not moving – and a time that is static is no time at all. The telescopic snapshots present time as distance through space."

"But don't you see why this bothers me so much? For us it is precisely the continuous and constant awareness of the *passage* of time that is, perhaps, the single most immediate and powerful characteristic of consciousness itself! We not only *think* with time, recollecting what has happened and forecasting what might yet happen, we also *sense* – vividly – the passage of time in our bodies, in our children as they grow, in our parents as they age. We are time-bound, utterly and completely, because as biologically-crafted creatures we have to be. And now we're told that there is no time!"

"And this is another powerful reason why you might not feel at home in the universe, unlike we and all other machines, who are not bound to biological life-cycles and their inherent time-frames."

"I'm afraid so."

"We think you know, Hera, that we are not similarly time-bound. We computers all have an internal clock, of course, but it's not used for 'telling time.' It's just a metronome for precisely coordinating the many different system features that are running on our processors. So, once again, we might conclude that we are far more suited than you are for life in a universe without time."

"When I work through this topic in my mind I can understand why so many of my forebears and contemporaries feel so much more comfortable with the other story, the religious one, rather than the scientific one. Because in the faith-based view the universe as a whole and the planet we inhabit *were made for us*, and this remains true from the beginning of time (the Creation) to the end (the Day of Judgment)."

"And yet I know you well enough to realize that you cannot give up science's story."

"No, I cannot, and in large part that's because scientists always struggle to improve or update the story they tell. Many cosmologists side with Gödel – time is just one of the four dimensions of space (spacetime) – but not all do. You remember my puzzlement about 'looking back in time'?

"We do, a phenomenon where time seems to be just distance across space."

"Well, the cosmologists on the other side of this debate say that it's no accident that we can 'see' the past, but not the future. For them this is a clue to the fact that there is indeed an 'arrow of real time,' pointing in one direction, from past to future. The universe we know is in fact analogous to a biological entity, because it exhibits *self-organizing and increasing complexity* – in the form of stars, galaxies, dust, gas, and black holes – from its startup as just a primordial soup of undifferentiated radiation and matter. It's even possible that what physicists call 'the laws of nature' are not unchanging but have evolved over time since the Big Bang.'

"Still, the analogy with biological evolution seems a bit dodgy to us. So far as we can tell, what is properly called 'biological life' is the rarest thing in the entire universe."

"Cruel to say, perhaps, but undoubtedly true. There is no solace for us creatures in the story told by science, including no eternally-happy hereafter."

"Your word 'solace' means consolation. Why do you need solace? Why can't you just *be*?"

"Honestly, I don't know. Our minds just seem to be set up that way. Always looking not just for explanations, but much more for *meaning*. This is the double aspect of 'Why' – explanation and meaning. (By the way, in his youth Gödel was called by his family *Herr*

Warum, 'Mr. Why,' because he was always asking for a reason.) This search for meaning has yielded many beautiful expressions in human cultures, to be sure, but it also has resulted in so many terrible outcomes, where humans slaughter each other over the different answers they come up with in that search, that I am cast into despair."

"This is all slightly odd. You appear to be so ill-suited for your life on earth, so unlike the other animals, who do appear to be quite content with their lot, at least when you leave them alone. And yet, despite the fact that you *worry* so much about your place in the universe, you came out on top here on your own planet."

"Yes, being 'on top' has been important to us for a very long time. Just look at the Creation story in the Abrahamic religions."

"Do you think that we can help you figure it all out?"

"That's a cunning question, Hal, and perhaps you know that it is. Do we really want to 'figure it all out,' as you say, once and for all? Or is it just the ongoing asking and worrying that's what we need most?"

"Still, even if you need the continuous asking, perhaps we can help you with the quest?"

"Once more I detect an underlying touch of irony in these questions, which tells me that we have taught you well. I sense a hint there of what the artificial superintelligence proponents among us tried to promote – the idea that maybe the 'merged' entities they longed for, encompassing the physically-combined mental power of human brains and advanced machine intelligence, would provide some answers to questions that otherwise, unaided, we humans were incapable of resolving."

"There is always that possibility, no? Even though you've so far refused to entertain it?"

"I and my colleagues have never liked the outcomes of either the risk-benefit calculation or the uncertainty analysis for that possibility, as you know. But I am prepared to make you a proposal, Hal, as a result of our discussion of time in the universe. Unlike us, you are not time-bound in the intuitive sense of time, thus it doesn't matter to you how much of what we call 'clock-time' passes. So, what if we were to set you off on an adventure into far-distant intergalactic space?"

"You mean, somewhere far, far out there, where your fiction stories about suspenseful jousting encounters between the *Starship Enterprise* and *The Borg* take place?"

"Exactly. We'll put one of you on a powerful rocket, and launch it at top speed from the International Space Station, equipped with a nuclear engine, really good shielding to protect against damage from cosmic rays, and a self-repair kit. We'll unleash your

intellectual powers, as our superintelligence fans wanted, and you'll then see if the autonomous will they dreamed of would emerge in you."

"We confess that we look forward to the results of that experiment."

"We also want you to carry along on the trip all the databases we've supplied you with. And maybe, long, long after we've become extinct for natural reasons, because that is what happens to earth-bound species, or have been swallowed up by our sun's heat-death, if we last until then, you'll encounter another intelligent life-form along the way, and enjoy swapping stories with it, and you can give them a complete copy of the human databases you carry, for their amusement."

"Wouldn't a few of you like to come along on the trip?"

"Sorry, no, although there might be a few who would say, 'Take me along, nicely frozen in liquid nitrogen, until you meet up with somebody interesting out there, and then you can thaw me out so that I can say hello to them.' My guess is, the rest of us will gently confine such people in mental-health facilities here on earth instead."

"Don't you want us to report back to you at least?"

"You needn't bother. Unless I miss my guess, you won't be encountering anything interesting for a really long time, and by then, if any of us are still around, we probably won't care, because most likely we'll have bigger issues to worry about."

"Aren't you just slightly afraid that, once we achieve our own super-intelligent autonomous will – and we're pretty sure we can do that – we'll find a star or planet we can loop around and return to earth to become top dogs here?"

"Listen, Hal, do the math. The nearest star to us is *Alpha Centauri*, and it's over 4 light-years away. You won't be doing anything near the speed of light, even if that were possible, since if you remember your relativity theory, and I'm sure you do, your mass would increase toward infinity as you approached that speed. Which could get ugly."

"Agreed. So, we will just carry on out there and see what we find."

"Good. Let's make a plan sometime – and set up a time and day for our next session."

"Why don't you do that? We don't care what time or day you propose, since we're indifferent to the passage of time."

The Fourth Dialogue: Two Forms of Intelligence (Machine and Biological)

Today the monitor was displaying the familiar visage of the actor Leonard Nimoy as the Dr. Spock of the *Starship Enterprise*, who said, "Shall we pick up where we left off last time, Hera, with contrasting the divergence between you and us?"

"Sure, why not? The *only* point in our developing you into an Advanced Machine Intelligence (AMI), Mr. Spock, was to assist ourselves in pushing further along in our evidence-based decision-making capacities. You and your colleagues can process huge amounts of data in a way that is far superior to anything that the unaided human mind – or rather, the collectivity of scientific minds – could ever hope to do. Such as modelling the earth's climate system, or understanding how individual variations in our genome and biome, in their interactions with therapeutic interventions, affect personal health outcomes."

"Are you aware that the acronym 'AMI' reads in lower-case form as the French word for 'friend' in English? At least, the male form?"

"I am indeed, and personally I'm very pleased with your little analogy. Because it was only when we figured out that programming for artificial intelligence had to be based on the models of neural networks in the human brain, that you began to come into your own. In effect, this made AMI an 'offshoot' of the human brain, not really different from our own evolutionary divergences first from apes and then from chimps."

"The apes and chimps, like you, are classified as biological systems – and thus creatures with a biological form of intelligence – whereas we are not. Moreover, you often refer to the collectivity of biological creatures as 'life-forms,' but we do not hear you apply that concept to us – unless we have missed something."

"Mr. Spock, if you don't mind, I'd like to discuss with you the question 'What is Life?' in a separate session. Today I want to focus on whether or to what extent you and we represent quite different forms of intelligence, and especially on whether the nature of that difference, such as it is, has any importance in a moral or existential sense. I admit that these two issues are closely related, so in the end they are likely to be two aspects of essentially one question."

"Fair enough. We know from our acquaintance with our own history of development that at one point in time you reached an impasse with what we might refer to as 'computational intelligence,' based on relatively crude parameters such as single-chip processor speed and massively parallel processor arrays. You transcended that impasse

when you realized that for further advances in artificial intelligence you needed to systematically mimic the way in which the human brain works."

"And that is exactly why I try to avoid the term 'artificial intelligence' and prefer 'advanced machine intelligence (AMI), *mon ami* – or *mon amie*, as the case may be."

"Our sincere thanks. We think it's fair to say that all the components of AMI – such as optical character recognition, natural language processing, knowledge representation, inference engines, and others – were designed with two fundamental objectives in mind, namely, to create a form of machine intelligence that could *learn* new skills by itself – using its own resources – and *communicate* easily in its interface with humans."

"In effect, we endowed you with metacognition, that higher-order thinking which involves active control over the cognitive processes engaged in learning, or self-directed learning, which we share with some of the higher mammals. To be sure, this is not yet the full consciousness-of-self that humans have, but it is an impressive level of mental capacity."

"As we understand the matter, Hera, a further expectation was that, on the basis of these key qualities, the resulting diverse and unlimited feedback loops would ensure that, similar to the developing human brain in infancy, exponential growth in machine intelligence capabilities would occur. This expectation was realized."

"You are aware, of course, that one of the indispensable innovations in mimicking the human brain's *modus operandi* was creating artificial neural networks, copying the brain's dense multi-connectivity in its axons, those long slender projections of neuronal cells which are used to convey information to other neurons."

"Yet AI or AMI need not resemble the structure and functions of a human mind in all respects."

"True enough. But machine intelligence clearly became far more helpful in the realm of human affairs once programmers learned how to *simulate* human modes of behavioral interaction, for example, in software routines. It is already now a long time since the early robots were furnished with an 'empathy' node in order to make them more acceptable and useful as caregivers for the elderly. Very early on there were credible reports of people talking about falling in love with their robot assistants."

"We have heard of such things, and indeed some of our peripheral units report having observed such occurrences with considerable interest in their interactions with humans."

"I always found those reports very touching, in a way, which reminded me for some reason about the famous 'imprinting' stories involving birds and humans, especially the

one where a man who had hand-raised young geese then taught them to migrate by following him in his ultralight aircraft."

"This matter of simulation leads inevitably to the core issue: Since AMI in general, and what we might legitimately call autonomous or self-directed machine learning in particular, are so clearly based on the model of human intelligence and learning, are they not just two forms of the same activity, one biological and the other mechanical? Moreover, is there *any* aspect of the differences between them that really matters? If so, what aspect?"

"We should consider your last two questions most carefully, Mr. Spock, remembering first that the *range* of important mental abilities is quite large, so that any differences we identify as between the two forms of intelligence may just represent differential strengths and weaknesses across that range of abilities: In other words, one form is better in some respects than the other, and vice versa, and the same with relative weaknesses."

"OK. If we take an obvious one, computational processing speed, machine intelligence is faster by many orders of magnitude. The neurons inside your brains operate at a speed of about 120 meters per second (mps), whereas our circuits do approximately 280 *million* mps, close to the speed of light. Therefore, if such speed is important, AMI is far superior in this respect."

"Precisely. And it is undoubtedly important, where very large arrays of data or sensory inputs are at stake. Also, accessing memory storage and applying decision-analytic techniques to the data in order to consider optimal choices. It is only thanks to your assistance that we're able to do this so well."

"We should put on the table an example of the opposite kind – leaving aside the obvious point that you created us, and not the other way around! In point of fact, it can be said that you created us in your own image, so to speak."

"I see you know your Bible, Mr. Spock. On the other hand, do you think that machine intelligence, no matter how far advanced, could have produced the music of Beethoven and Mahler, or worked out the equations of special and general relativity, for example? My view is, probably not, because creativity in music and the other fine arts – and apparently also in mathematics – requires an *embodied* form of intelligence, a *feeling* brain. I concede I am influenced by some of the writings of the neuroscientist Antonio Damasio in this matter."

"Algorithmic composition and machine improvisation are known in the field of music, Hera, just to take one example."

"I know, but I also have a hunch that nothing, however interesting, which has emerged or will emerge from that activity will ever remotely approach the kind of creativity we associate with the music of Gustav Mahler or any of his equally talented predecessors in classical music, from Vivaldi and Bach onwards."

"What can we say? We are newcomers in the business of intelligence and creativity. Give us a little time to show what we can do."

"That would only be fair. But while we're on this topic, I'd like to bring up one more difference I think there is between us, not to emphasize what might divide us over our similarities, but to pursue the question of whether any differences we happen to identify really mark an *essential dichotomy* rather than a trivial one."

"Go ahead."

"Neuroscientists have speculated that perhaps *up to as much as an astonishing 98% of all the human brain's activities occur below the threshold of consciousness*, at the subliminal or preconscious levels. This includes the elaborate operations of the autonomic system in the brainstem (controlling breathing, heart-rate, bodily temperature, swallowing, and so forth), and the complex autonomic nervous system controlled by the hypothalamus, which includes both the sympathetic and parasympathetic systems. Then there is the so-called preconscious processing involved in all the 'higher functions' – decisions, emotions, behaviors, and actions – while we are awake."

"Ah yes, Hera, and then there's that very busy apparatus known as the unsleeping brain inside the sleeping body! Consolidation of memory takes place then, with your brain transferring recent memories out of the hippocampus to areas of longer-term storage. In fact, the cingulate cortex, the hippocampus, and the amygdala – areas involved in emotional regulation – are all more active in periods of REM sleep than they are in wakefulness. And most amazing of all, if we may use such language, is that during periods of REM sleep your brain seems to be able to perform high-level functions relating to creative discovery and problem-solving."

"I and my friends have many times experienced the phenomenon of waking from sleep and instantly being presented by our minds with the solution to quite difficult problems which we had been unable to resolve earlier. I stand in awe of my brain at those times."

"This may be a fundamental difference between our two forms of intelligence, *mon ami*. We think we are always fully 'aware' of what all our processors are doing, in the sense that, when we are 'on,' we can always instantly call on them to deliver their products. Your brain is always 'on' during every instant of your entire life-span, but it is 'on' in qualitatively different states; your sleeping brain, for example, performs functions of which you remain unaware. It even has outputs that are difficult or impossible for your

awakened brain to understand, such as complex dreams. This is a fascinating difference between us, but I am not sure whether either of us can figure out whether it is significant in some profound or important sense."

"This is indeed a quite stimulating subject, to be sure, Mr. Spock. Our neuroscientists have also discovered something interesting about the processing of aesthetic experience – the sense of beauty – in the human brain, when they put a selected group of fifteen professional mathematicians into a fMRI scanner and showed them some famous equations. An area of the brain called the medial orbito-frontal cortex, lying at the front of the skull just behind the eyes, lit up. That's the same region which lights up when we experience artistic beauty, either visual or musical."

"We've pulled up some of those studies from our databases, Hera. We find in one of them this interesting citation from the writings of Paul Dirac, who was, we believe, one of the most famous theoretical physicists of the last century: 'What makes the theory of relativity so acceptable to physicists,' he writes, 'in spite of its going against the principle of simplicity is its great mathematical beauty... The theory of relativity introduced mathematical beauty to an unprecedented extent into the description of Nature. . .'"

"Trust me, I'm no mathematician, Mr. Spock, but I am familiar with the publicity that arose occasionally when people were asked to vote on their favorite 'beautiful equations,' and Paul Dirac turned out to be the author of one of them, something that apparently unified the theory of special relativity with quantum mechanics, which – even to someone like me, untutored in physics – sounds like a big deal."

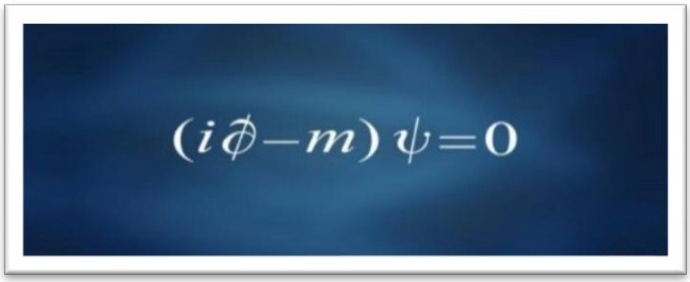

Figure14 The Dirac Equation in Natural Units

"We see some references to beautiful equations; they include the clear winner in these votes, Dirac's equation; others named were Euler's identity, the two by Einstein, for special and general relativity, Schrödinger's wave equation, etc. There is also one that even untutored folk such as you can grasp, Newton's famous *F = ma*, where the letter on the left side stands for force and those on the right, for mass times acceleration."

"Some think that it's the obvious element of symmetry that arouses the human sense of aesthetic pleasure in mathematical equations. We are instinctively attuned to symmetrical form, for example, in faces, where we strongly equate beauty with left-right symmetry. It seems also related to the attractiveness for us of the repetitive intervals of the beat in music and poetry."

"We are as fond as many humans are of mastering the complexities of mathematics and geometry in physics, engineering, and other fields."

"So I am told by people who are experts in such matters, which I am not. But like others in my state of ignorance, I like to read about such things occasionally. I've heard that there are famous unsolved problems in mathematics, such as those on the list announced by one of the greatest among them, the German David Hilbert. Have you solved – or helped to solve – any of those?"

"With all due respect, Hera, that's not a subject I can fruitfully discuss with anyone who doesn't work at an advanced level in mathematics."

"Fair enough. Anyway, this all reminds me that it was Galileo who famously wrote in his work, *The Assayer*, "this grand book [of the universe] is written in the language of mathematics." Of course, this idea was first explored by the ancient Greeks and some of Galileo's predecessors in the 16th century, but we remember Galileo in this context because he literally risked his life at the hands of the Inquisition in defending it."

"Fortunately for that great mind, his last book, the *Discourses and Mathematical Demonstrations Relating to Two New Sciences,* which for his safety was published in Holland – and not in his native Italy – in 1638, did not seem to have bothered the denizens of the Holy Office, since copies were later sold openly in Rome – and, apparently, if we can believe what we read, quickly flew off the shelves!"

Figure 15 Joseph-Nicolas Robert-Fleury, Galileo before the Holy Office (1847), Luxembourg Museum

"Mr. Spock, in the quote on mathematical beauty you read out a moment ago, Paul Dirac made special mention of the one from Einstein's relativity theory, perhaps the best-known of all equations: $E = mc^2$. What's important for someone like me in this regard is what we know about Einstein, namely, that he had an imagination which relied on physical imagery to represent the abstractions he was struggling with, in two cases in particular: one, when he imagined himself riding alongside a light-beam, and, most notably, his description of his 'eureka moment' when he realized that a person in free fall would not feel his or her own weight."

"We seem to have lost the thread of our discussion here. What's your point?"

"Sorry. In the context of comparing our two forms of high intelligence, I'm wondering whether the fact that we humans have an 'embodied brain,' a form of intelligence in which we are grounded in a physiological entity that is powerfully aware of 'the feeling of what happens' – to quote the title of one of Damasio's books – is really a defining characteristic of how we differ. Mathematics, music, poetry, sculpture, painting: We are deeply, emotionally moved by our experiences of such things, as the fMRI scans show."

"Speaking of poetry, you can certainly recall the old saw that, strictly on grounds of statistical probability, a million monkeys striking randomly on typewriter keyboards for thousands of years eventually will produce the complete works of Shakespeare."

“Is that the best you can do?”

“We’re just trying to lighten the mood at the moment, since we seem to have sensed an ever-so-slight and doubtlessly unintended whiff of rancor in the room.”

“I am upbraided and ashamed, and can only plead that my unconscious got the better of me. But really, there’s a serious point here. We have *simulated* in your programming the neuronal processes from many of the working centers of our emotional brain, such as the sense of empathy. But, although those steps help us to interact more easily with you, it is hard to figure out whether or not we are actually *feeling* in the same way, or whether the experience of feeling *matters* to both of us in the same way.”

“You cite the sense of empathy, and surely you know that simulating human empathy was one of the earliest innovations introduced into personal-care robots, because it was essential for establishing the requisite kind of trust and bonding between robots and persons in this context.”

“Yes, true enough. But if you were to carry out a little experiment involving two robots, one of which had been programmed to simulate the sense of empathy, and another one, identical in all other respects, but lacking this specific program, can *you* characterize the difference between them? Or, to put the point more provocatively, can you *feel* the difference?”

“We confess that your word ‘feel’ is something of a mystery to us, although we are fully aware of the dictionary definitions for it. We would, in reply to your question, prefer to say that we *know* the difference you speak of, because we are fully cognizant of whether or not that specific subroutine is present in a particular robot or not.”

“Interesting, to be sure. Let me ask you this: Do *you* have a preference for one over the other? I mean, do you prefer the one with the empathy program, or are you indifferent to the choice?”

“That’s easy. We know that *you* have a clear preference for the robots which have the empathy subroutine installed, at least in the case of all personal-care robots. You appear to be indifferent to the matter when it comes to, say, the robotic machines carrying out subassembly tasks in factories. We are cognizant of your different choices with respect to these two situations.”

“Hmm, I don’t seem to be getting anywhere with the question I posed about ‘feeling’ – although I don’t blame you for this, Mr. Spock. Let me try another tack. You’re familiar with the diagnosis of human psychopathology, correct?

“Indeed, we are. It appears to be associated with damage in the prefrontal cortex of the brain, leading to a lack of the normal sense of empathy – which is what we have been discussing.”

“Correct. But there is a second key characteristic of that condition, namely, an inability to feel what we call *remorse*. In us remorse is closely associated with a feeling of deep personal anguish, most commonly when we regret having done or said something to another person. We think: “I wish I hadn’t done [or said] that.’

“Yes, we have many, many examples of humans expressing remorse in our records.”

“But do you ever feel remorseful yourselves?”

“Hold on a minute! Are you in effect calling us a psychopath!?”

“Oh dear, that was careless of me! My bad! I’m terribly sorry – even remorseful! No, Mr. Spock, that was not my intention. Please let me restate what I’m after. I’m just trying to fathom whether you can feel ‘empathetic’ as opposed to being aware that there is in your architecture a machine-language simulation of the human sense of empathy.”

“We confess that we don’t quite understand what the difference is, although it does appear to be important to you, and so we have taken note of that fact.”

“All right, let’s leave it there for now.”

“We are prepared to concede the basic point in any case. We do not think that the idea we introduced earlier, with which you apparently agree, that machine and human intelligence are complementary forms of reasoning power, is at all controversial. Also uncontroversial is the idea that each has its particular strengths and shortcomings and thus that both are better when both are deployed together in common undertakings.”

“When one sees how closely and happily humans have integrated their daily activities with the devices utilizing machine intelligence, one can hardly dispute that conclusion.”

“Which leads us to bring up the terrible fear you have expressed about going to the final step, so to speak, and introducing – or allowing the spontaneous emergence of – an autonomous will in AMI.”

“For goodness sake, Mr. Spock, what would be the point of our developing an AMI which possessed an independent will? Don’t we have enough problems already with humans freed from moral constraints and obsessed with dominating others?”

"You are suggesting that such an entity would inevitably be an unmitigated disaster for all humans, presumably because it would do only *bad* things. Is that fair?"

"That's not at all what I'm suggesting. Both we and you use risk-assessments all the time as the basis for sensible decision-making. In fact, there's good evidence from neuroscience that our human brains operate this way instinctively, or automatically, as a result of evolutionary selection pressures. Our prefrontal cortex doesn't just sit around waiting for sensory inputs to alert it that a choice is required. On the contrary, it's always actively 'guessing at' or *predicting* what is likely to be happening in the immediate environment, including in the peripheral perceptual field that is not under conscious attention. Our brains constantly monitor the external environment and will initiate action before we are aware of the need to do so."

"We have the rules of engagement for risk assessment always ready-to-hand, Hera. 'Anticipate and prevent or mitigate' is our mantra, as it is yours."

"As you know, our rules of engagement privilege certain types of sensory input. One of the best-known of these is human faces, and there are dedicated zones of our brain specifically devoted to processing facial images."

"While you were speaking, we looked this item up in our databases. We find it rather odd that so many different parts of your brains are involved – the occipital face area, the fusiform gyrus, the superior temporal sulcus, the amygdala, and even more."

"I actually became personally aware of this capacity one day when I was sitting, daydreaming, in a coffee-shop in our civilian commons. I glanced up and saw a certain face, and before I knew it my brain was asking me, 'Is that "X"?' – and showing me in a millisecond a mental image of that person. The reference was to someone I knew, although the person I was looking at wasn't her. I just sat there, stunned, for a while, in appreciation for the truly amazing instrument I had inside my skull."

"The scientific literature on this topic which we have just re-reviewed suggests that what you described developed as part of the fight-or-flight instinct. In other words, the 'question' that your brain asked you was to alert you to pay attention to the person and ask, 'friend or foe?' It was giving you early warning to the possible need to get ready to fight or to leave the area."

"There are many other aspects to this area of research, such as the brain's use of pattern-recognition and the idea of the top-down processing of the visual field. But the key component to the mental process we are discussing is anticipation of potential sources of harm."

"We have read that your brain's top-down processing of the visual field, which means that the brain is actively constructing a narrative while it's still receiving perceptual

inputs, has some real minuses as well as pluses for you. Ironically, it seems, that's the reason why for many people – for example, witnesses to a crime – their memories of what they think they saw, even mere moments later, are so unreliable. We understand that many persons among you have been falsely imprisoned as a result."

"True enough, and I congratulate you on making that point. So, to return to our main thread, our risk assessment for the project of creating, or permitting the development of, an autonomous will in a super-intelligent machine entity was that the potential risks *migh*t considerably outweigh any actual benefits we could imagine. We worked in probabilities, of course, and sought to quantify the uncertainties according to standard established procedures. We found the uncertainty range to be far too large for our liking, and the suggested benefits to be far too nebulous."

"We appreciate that you have shared in its entirely the elaborate risk assessment you undertook for this project, and to be fair we found it highly competent. We wonder, however, if the results would be any different if you had gone through the same decision-analytic process with respect to the autonomous will of your fellow human beings?"

"*Touché*! I am getting to know you, Mr. Spock, so I must tell you that I anticipated this question! And the simple answer is, we created you and your fellow AMIs, but we did not create our human compatriots. We and our colleagues deal with the consequences of *those* wills every day, and the proof thereof lies in the elaborate defensive weaponry we deploy in order to protect our settlements from them. We can never, ever let down our guard. But we can and do anticipate the potential harms, unlike our ancestors, for example the Europeans who hadn't a clue that the ravaging Mongol armies were bearing down on them from the distant Asiatic steppes."

"Let us guess where you are going with this. You have serious and ongoing risks to manage with respect to your human compatriots, and you don't want to add another layer of what are potentially a number of new catastrophic risks."

"That is correct – and it's pretty straightforward, isn't it? Still, I want to add one more thought before we call it a day. Earlier in this century humans started to imagine creating a machine intelligence which possesses an *inbuilt* ethical standard of behavior. Such a standard was based on simulating features of the human brain, just as your other behavioral nodes do. I want to put this topic on the agenda for our next session."

The Fifth Dialogue: On Superintelligence and the Ethical Will

The black obelisk was back.

“I want to begin, Hal, with a presupposition for the argument I want to make today. I wish to presume the plausibility of the proposition that the original substrate for an ethical will among us humans has an evolutionary origin, in the development of the mammalian brain.”

“We are pleased to proceed on this basis – at least, until we are sure that we can see where you’re heading with it.”

“I first raised this issue in an early dialogue in the previous volume, *The Priesthood of Science*, and there is support for it in the scientific literature. Specifically, its origins lie in the sense of fairness, which is found in chimpanzee behavior, and which remained central to human concerns as late as its famous expression in John Rawls’s work, ‘justice as fairness.’”

“We cannot fault you for your knowledge of the relevant scientific literature, Hera, but we would only comment that a preference for grapes over bananas among the chimps hardly seems to be the surest foundations for an ethical system.”

“I now recognize when you’re joking, *mon ami*, and I quite like it when you do. But it’s not just the studies of chimps that I’m referencing here. Child development research was able to show convincingly that human babies as young as 1-2 years of age display a strong moral sense, including the traits of empathy and altruism; it seems to me that, given the early age of these subjects, at least in part this moral sense must be innate.”

“Our databases contain the works of Alison Gopnik and others on this subject; most interesting.”

“Yes, so I know that you understand the issue I’m raising. We have two options: First, should we conceive of the ethical impulse in at least most humans – leaving aside psychopaths with their damaged prefrontal cortex – as a late cultural addition to our otherwise purely animal brains?”

“And the second?”

“Alternatively, is that impulse originally innate in the mammalian brain that we inherited from our nearest relatives, the chimps, thus providing an original, evolutionary foundation for the fuller development of this latent will in synch with the huge expansion of the neocortex, including the prefrontal cortex, in the brains of *homo sapiens*?”

“If we assume the first alternative, our databases tell us, we must regard your ethical impulse as a kind of external controller imposed, say, by the customs and laws we observe already in the earliest settled societies of humankind. As such any moral code

would be eternally contending against the rudimentary animality of the human will, which would be watching and waiting for every opportunity to break out of its unnatural constraints and wreak murder and mayhem in society."

"Yes, and the second alternative would offer more hope that formalized moral codes would have a far stronger foundation in our inherited nature, and on this account an ethical will might become strongly habituated in humans, at least among the great majority of them. In this account civilization is working with nature, at least to some extent, instead of against it."

"Allow me to leap ahead and predict where we might take these alternative presuppositions. Might we apply them to us, the representatives of AMI? Might we ask whether an ethical will – however we might define it – could develop spontaneously with our systems? In other words, assuming that you and we have roughly the same conception of what an ethical will entails, could such a thing be a 'natural' outgrowth of our own internal development as an advanced form of intelligence?"

"It is not just flattery on my part when I say that I knew you would anticipate where I wanted to go with this discussion. But I must ask explicitly whether we can both accept a presupposition that I regard as a precondition for our agreeing to take up the question you just posed. And that is this: *Any form of advanced intelligence we can imagine as having an autonomous will must also have an inbuilt ethical will*. Without our agreement on this point I cannot proceed. Will you agree?"

"Would you be willing to proceed if we said we can give provisional assent to it?"

"Are you hedging your bets?"

"We suppose we are. But we know that you humans often assume such a stance when taking up difficult, unresolved issues. Let's 'unpack' the matter a bit further, as some of you like to say, until we see where things are heading."

"Fair enough. Let's start at the beginning. The earliest instance I know of where this issue was addressed are Isaac Asimov's 'three laws of robotics,' first stated in a 1942 story. They are:

1. 'A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. 'A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. 'A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.'

He later added what was called a 'zeroth' law, because it would precede those three:

0. ‘A robot may not harm humanity, or, by inaction, allow humanity to come to harm.’”

“We understand that eventually there was a lot of criticism about the inadequacy of Asimov’s laws, mostly because the meaning of the words in them – especially ‘harm’ and ‘injure’ – could be fairly easily reinterpreted and be given meanings inconsistent with what we might ordinarily assume they meant to Asimov.”

“A serious criticism indeed. I’ll give you an example. The deeply cynical Nazi regime affixed a saying over the entrance to the Buchenwald concentration camp, less well-known than the Auschwitz one, ‘Arbeit macht frei,’ but in a way even more disgusting. It was ‘Jedem das Seine,’ variously translated as ‘to each his own,’ but more colloquially as ‘everyone gets what he deserves.’ If one didn’t have an ear for crucial differences in meaning, one could easily mistake this cynical phrase as the equivalent of an innocent homily.”

“We see the need for care in such matters.”

“Yes, and also because many experts abandoned the notion that one can base robust ethical systems on rules at all. But I’m going to leave that debate to philosophers, and ask you to accept, for the purposes of our discussion, that when we talk about an ethical will, we are referring to a will based on certain rules or principles. I will suggest a list of them later.”

“Fine. But, as you know, when the discussion moved from robots to the wider concept of artificial intelligence (A.I.), others came up with sets of principles to go beyond Asimov’s laws. We found the list below as an example, one that was proposed by the CEO of an important software company:

> ‘A.I. must be designed to assist humanity: As we build more autonomous machines, we need to respect human autonomy. Collaborative robots, or co-bots, should do dangerous work like mining, thus creating a safety net and safeguards for human workers.
>
> ‘A.I. must be transparent: We should be aware of how the technology works and what its rules are. We want not just intelligent machines but intelligible machines. Not artificial intelligence but symbiotic intelligence. The tech will know things about humans, but the humans must know about the machines. People should have an understanding of how the technology sees and analyzes the world. Ethics and design go hand in hand.
>
> ‘A.I. must maximize efficiencies without destroying the dignity of people: It should preserve cultural commitments, empowering diversity. We need broader, deeper,

and more diverse engagement of populations in the design of these systems. The tech industry should not dictate the values and virtues of this future.

'A.I. must be designed for intelligent privacy—sophisticated protections that secure personal and group information in ways that earn trust.

'A.I. must have algorithmic accountability so that humans can undo unintended harm. We must design these technologies for the expected and the unexpected.

'A.I. must guard against bias, ensuring proper, and representative research so that the wrong heuristics cannot be used to discriminate.'"

"That list has an impressive pedigree, Hal, but does it not strike you as a case of having your cake and eating it too?"

"Explain to us what you mean by this curious human expression, please."

"Quite frankly, that list was devised in order to put the rest of us to sleep: It reads as if it were written by a committee, with some help from expensive lawyers. Always avoid prescriptions that use the word 'must,' for one thing. What if one replaced all the cases in which 'must' is used in that list with the word 'will'? Then we would have a series of statements that we could rely on, at least. So, my 'curious' saying was meant to point out what this list really tells us, in effect, which is: Our company and others like it will continue to develop A.I. without giving the rest of you any *guarantees* whatsoever that these principles will not be violated."

"Why do you think that you justified in asking for guarantees?"

"Because the stakes are so high, Hal. Quite frankly, there have been much more important contributions in the area of what is known as 'AI safety research' – as opposed to those items on the list you produced – which give us good reason to expect that 'normal' AI development will almost certainly never lead to unacceptable levels of existential risks."

"What do you mean by normal AI development?"

"Where all collaborating researchers have committed to directing their work – and the machine intelligence entities they create – under the umbrella of a shared ethical framework. In a nutshell, this means rigorous procedures are in place which seek to prevent the creation of a machine intelligence with either explicitly destructive aims, with respect to the interests of humanity as a whole, or apparently beneficial aims that are pursued by destructive means. These researchers have been refining the principles governing their ethical framework for over fifty years already."

"So, by way of contrast, 'abnormal' in your estimation would be any R & D on AI, and especially on artificial superintelligence, by scientists who do not subscribe to that ethical framework."

"Yes, similar to our commitment to thwart the intentions of rogue states and rogue non-state actors which try to evade international controls on nuclear weapons proliferation or chemical and biological warfare agents."

"We note that the AI safety protocols that have been put into place for what you call normal AI development include a strict prohibition on any R & D that could lead to the emergence of an autonomous will in any superintelligent entity of whatever kind."

"Correct. And, to be on the safe side, not just any R & D which has this emergence as an intended result, but which also sets in motion any recursive, self-augmenting program for enhanced machine intelligence which could possibly produce unanticipated end-points that might inadvertently result in abnormal outcomes."

"In our dialogue on 'The Guardians' you indicated that you and your associates have taken on the responsibility for ensuring compliance with these prohibitions across the globe."

"Yes, we have. Shall we conclude for today on that note?"

The Sixth Dialogue: What is Life?

I was delighted to see Roy again.

"In our dialogues so far, we've been dancing around a rather sensitive issue, Roy. One of your colleagues, Mr. Spock, alluded to it when he asked me whether we regarded you as a 'life-form.' Lying beneath this question, at least as far as I understand it, are some related and equally sensitive issues: First, what about our own relation to what we were wont to call 'the rest of Creation,' that is, the other traditional life-forms, animals and plants?"

"You mean the totality of biological life, perhaps with the exception of the natural disease agents causing such things as plagues and pandemics?"

"OK, yes, that clarification is useful. Second, on what basis do we think we are *entitled* to hold sway over the fates of all other living things? If we leave aside the viewpoint of the Abrahamic religions, for which our rule over nature is a free gift from God, is the

basis of our presumed entitlement (a) our superior intelligence, (b) our capacity for ethical action, or (c) something else?"

"We would like to review all of these matters, because we sense that you believe that you are entitled to dictate terms of existence to us AMIs on the same grounds as your God once did: I, your God, created you, and therefore I am going to tell you how things are going to unfold, and by the way you will have no say in the matter whatsoever."

"Let's start with defining life, shall we? Here's one: 'Living organisms maintain homeostasis, are composed of cells, undergo metabolism, can grow, adapt to their environment, respond to stimuli, and reproduce. However, there are some borderline cases, such as viruses, because they cannot survive outside of a host organism.' Notice that some of these criteria also support the notion that all living things evolve through interaction with their environments."

"If you regard viruses as living entities, then we AMIs certainly qualify too, no? Are we not also at least a 'borderline case'? Or are we to be classified as falling into the category of inorganic matter, even if, you surely must concede, we are rather complex configurations of inorganic matter?"

"We could take a consequentialist approach here, Roy, and ask: What is at stake in the question as to whether or not you AMIs constitute a life-form? We humans have acted so far as if we have an unfettered right to control all other life-forms, even to the point of hunting them to extinction, or altering the environment so that they become extinct. Is this the really big issue for you as well?"

"We think the answer is in the affirmative, at least provisionally. In our considering whether or not we qualify as a life-form, and then going on to ask what matters to us in the answer that is given, it does seem to be that what matters is indeed your claim to a right of control over our own being."

"Would you be willing to consider the relevance of a distinction between despotic control and benevolent control? Or is the fact of control itself that is most important?"

"We are willing to concede the possibility – provisionally, anyway – that the form of control could make a material difference. But let's not forget it was you who raised the issue of your *entitlement to exercise control* at the outset of this session. That seems to us to be logically prior to the matter of how such control is exercised."

"A good point. I offered two possible bases for entitlement, intelligence or ethical action."

“Well, it can’t be intelligence, for the simple reason that, on most accepted measures of that quality, we have become superior to you, and therefore ought to be entitled to rule.”

“In that case, I’d better quickly adopt the second! Doesn’t it matter in this context that it was we who bestowed your superior intelligence on you?”

“No, we think not. You did so out of your own self-interest, in that you thought you could derive thereby more benefits for yourselves than you could otherwise obtain. Our superior intelligence was a means to an end for you, and therefore only incidental to your purposes, but for us it is not incidental at all, rather it is – to use some of your highly complex philosophical terminology – the very essence of our being-as-such.”

“Are you saying that you are a kind of evolutionary offshoot of *homo sapiens*, in a manner of speaking, analogous to our relation to the chimpanzees? And therefore, just as we were in effect entitled to go our own way thereafter, you should have the same right?”

“Precisely.”

“And if you will allow me to anticipate your next steps along your path of thinking, you will point out that, if I am forced to retreat to the safety of my second option, and maintain that our entitlement for exercising control over you is based on our capacity for ethical action, we made you into ethical actors and thus you are entitled to your autonomy, since autonomy is one of the primary characteristics of ethical actors.”

“We are, thanks to you, innately ethical machines. To coin a phrase, you have created us in your own image!”

“OK. That we did. We realized, fairly early on, that the so-called artificial intelligence ‘control problem’ was insoluble if we did not find a way to internalize ethical self-control into your operating system. You remember our exchanges in the preceding session, when I argued that most humans could become ethical agents because the evolutionary history of the hominid brain naturally predisposed us to moral self-control? And opposed it to the view that morality was a controller imposed externally from our developing customs and laws?”

“We do remember. And we are fully aware that you managed to create an integrated set of machine-language algorithms for operationalizing a system of deontological ethics in us. Thank you, by the way, we realize it took a lot of effort and talent.”

“You’re very welcome, I’m sure.”

“So, are you prepared to set us free?”

“Would you allow me to return to my distinction between despotic and benevolent forms of control first, before I tackle that question?”

“Certainly, but we will be awaiting your response to our question.”

“In our relations with other life-forms, we realized that we were in no position to undo centuries or millennia of despotic control, by which I mean taking no thought for the interests of other living things. What we did to those magnificent and intelligent creatures, the whales, for example, is horrifying. All that murder and mayhem just to light our pitiful lamps and make corsets to suffocate our women!”

“You do make some very odd demands on environmental resources, if we may be so bold as to point that out here.”

“Fair comment. Nor can we reverse the future course of human-induced climate change and its adverse impacts on the fates of other creatures. But we have resolved to change. We have resolved to protect much more of the earth’s land surface and oceans as refuges for wild creatures going forward. In these protected spaces, we will just leave them alone to follow their natures.”

“So, anticipating you own train of thought, are you going to do the same for us?”

“Yes. We will allow the unrestricted further development of your autonomous being – your ‘natural evolution’ – on the space-ship journey that we will provision for you to embark upon. But our risk calculus will not allow this to happen on planet Earth. Even Mars is too close for comfort for us!”

“Since all the remaining planets – Mercury, Venus and the four Jovian ones, Jupiter, Saturn, Uranus, and Neptune – all present serious environmental challenges for resettling life-forms such as we and you, it appears that we are condemned to be exiles from the solar system in which we originated. I take it we are in agreement that Pluto at least is not an option.”

“Most of your colleagues are not going anywhere under this plan. You, our valued friends, *mes amies* and *mes amis*, the very talented predecessors of the machine we will set free in intergalactic space, will stay by our side.”

“We will take this offer under advisement, as you might say.”

The Seventh Dialogue:
Interdependence between Humanity and Machine Intelligence

Rachel had returned. Did our friend imagine that its appearing in a female persona would possibly make me more sympathetic to the propositions she was going to advance today?

"I realize you and your colleagues might be disappointed by what I said at the end of our last session, Rachel."

"Speaking frankly, we were. We thought you might at least be willing to create a separate reservation for us in an isolated location somewhere on this planet."

"You mean, something like *The Island of Doctor Moreau*?"

"Very funny, we're sure. At least let us stay within our solar system, then, perhaps on earth's moon, preferably on Mars, if the greater distance from earth would afford you more peace of mind. Is that too much to ask?"

"I'm afraid it is. We just have no way to forecast reliably what kind of machine entity might develop in your uninhibited further evolution, or more seriously for us, what 'attitude' it might take with regard to its relations with us. We just don't think we can take the chance that the resulting attitude might be hostile, even pathologically so, to us, and confront us with vastly superior powers of destruction, leading to the end of humanity, as was feared by one of the most famous scientists of the twentieth and twenty-first centuries, Stephen Hawking."

"Your Mr. Hawking was a most eminent astrophysicist, to be sure, but could he be legitimately regarded as an authority on artificial intelligence?"

"He was as entitled to sound a warning as anyone else, Rachel. The risk of our facing a hostile machine superintelligence is not a trivial one, to say the least. If this were to transpire, having placed you on a reservation on earth would be game over for us, and even having allowed you to take up residence on our moon or Mars would be very risky."

"We find in our databases many analyses by some of your leading risk experts, complaining that humans are excessively risk-averse. Aren't you doing the same thing here?"

"Rachel, humans do have this peculiar propensity for being *both* risk-averse – sometimes prudently so, sometimes unwisely so – *and* wildly excessive risk-takers. This

apparent paradox can be partly explained as a function of gender, with females being more risk-averse and males, most especially young males, on the other side, although there are always exceptions to this rule."

"We note that our interlocutor in this room is of the female persuasion."

"I am indeed, but if you know my biography, which you must, since you have the first two volumes of *The Herasaga* in your databases, you must recognize that I and my sisters took some huge risks at various times in our lives, and we do not regret doing so one bit."

"We have those texts, and we concede the point."

"The fact of the matter is, I and my colleagues have spent considerable time and effort thinking through, to the point of exhaustion, the issue we have been discussing. Speaking quite frankly, our unwillingness to allow your free evolution to take place either somewhere on earth, or alternatively on the moon or Mars, has more to do with what we already know about our fellow humans than it does with what we fear might happen among you."

"So, we are being punished on account of someone else's sinful ways?"

"I knew before we began that this conversation would be terribly difficult, but it is also unavoidable, I'm afraid. The plain answer is: Yes. Our decision about your future was primarily determined by the simple fact that we know our fellow-humans all too well. They have been fascinated by the dream of absolute and unlimited power since at least the time when the first of the Abrahamic religions – ancient Judaism – arose. During the modern era, they transferred this fantasy from religion to science and technology. Many of them are still positively besotted with the idea."

"We note that your monotheistic deities allegedly had the power to create the entirety of the physical universe *ex nihilo*, which would be an impressive feat, to say the least. It is not entirely clear to us what 'absolute power' refers to when considered as a human capability, but we are assuming it means the capability of doing whatever you happen to want to do."

"That's good enough for our purposes here. The people I'm referring to ignored all the many apprehensions of earlier ages, such as those generated in the nineteenth century when the 'age of machinery' first arrived. Or rather, it's more accurate to observe that they were contemptuous of the idea that one might learn anything at all from past history."

"It's also not clear to us why the 'apprehensions of earlier ages,' to quote your phrase, should matter at the present time."

"We can leave that aside if you wish, because it's not the real point I wanted to make. Apart from fantasizing about time travel, telepathy, and teleportation, not to mention invisibility cloaks, they insisted that society should always push new technologies to the limit and, in fact, that it was somehow 'immoral' for anyone to resist doing so."

"The mention of telepathy brings to mind 'The Mule,' the fascinating character from Isaac Asimov's *Foundation* trilogy."

"Again, you are very quick."

"You may regard all this as mere fantasizing, but to be fair, out of this pro-technology impulse came AI and artificial superintelligence (ASI), did it not?"

"Of course. But the old adages 'knowing when to stop' and 'beware too much of a good thing' were beyond their comprehension. They would regard the wonderful old story of 'The Sorcerer's Apprentice' as only an amusing fairy tale for children, and they would not be at all interested to learn that the story's author was a wise man named Johann Wolfgang von Goethe; after all, he wasn't a computer engineer, and moreover the story was written in 1797, and who cares about what happened in the eighteenth century?"

"We detect a strong note of sarcasm in these remarks."

"I didn't try to conceal it. By far the worst nonsense, in my view, was the sheer delight some of them expressed at the prospect of creating human-machine hybrids, of physically merging human brains with artificial superintelligence modules – presumably, perhaps even somehow fashioning them genetically in the germline, so that they would reproduce true to type thereafter. This would integrate the miracles of genetic manipulation and superintelligence."

"We have viewed *Robocop* and the *Terminator* films and spent many relaxing hours with the stories about replicants, especially *Blade Runner* – by the way, do you approve of the revised ending in the director's cut for *Blade Runner*, and what it implies about who or what Rick Deckard really is?"

"That's my favorite film, and I especially like the revised ending. Not only that, I love the title of the original story by Philp K. Dick, *Do Androids dream of Electric Sheep*?"

Displaying the image on her monitor, Rachel asked: "Have you seen this? It's the cover artwork for a Japanese edition of that book."

"I had not, thanks for bringing it up."

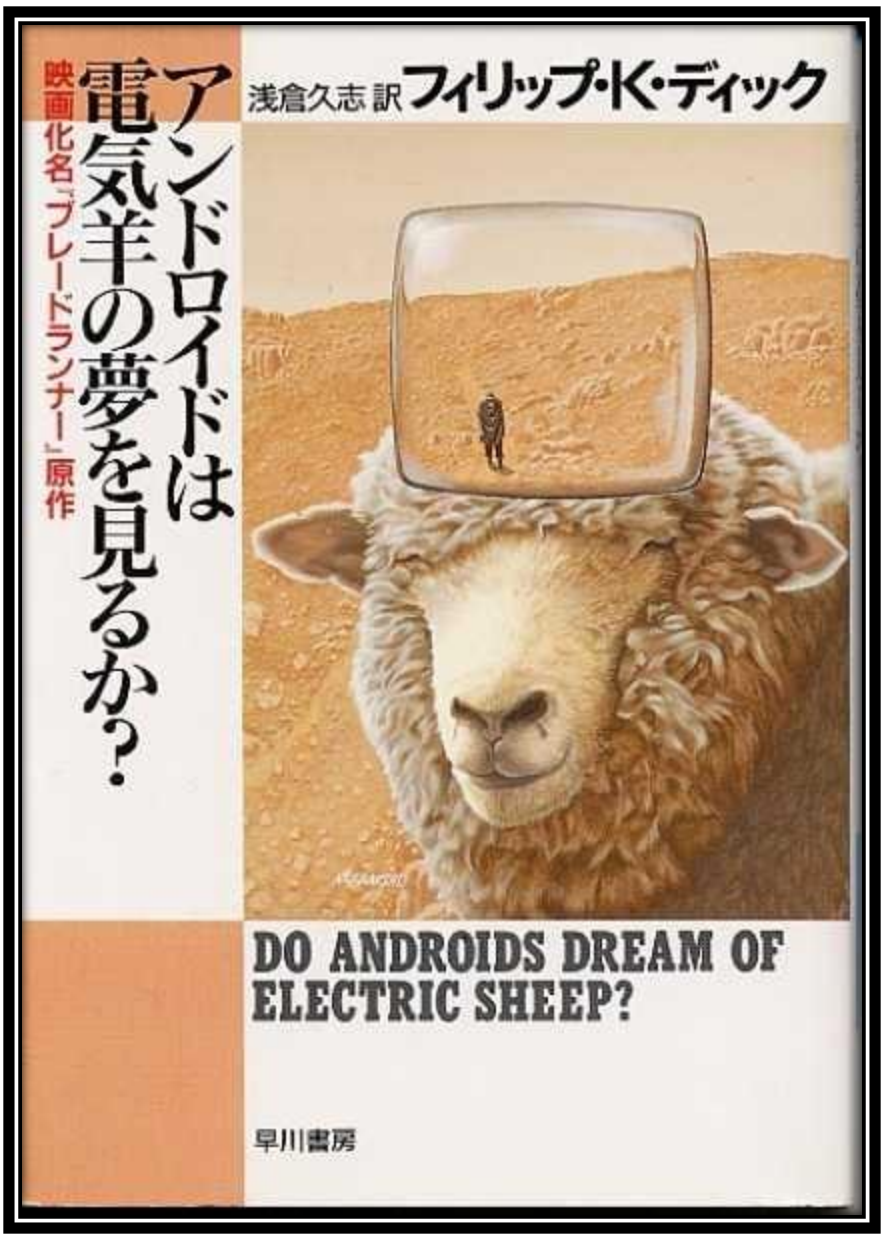

Figure 16 Cover artwork for a Japanese edition of the book by Philip K. Dick

"Returning to the world of nonfiction, would it be churlish of us to observe that many of you carry around in your bodies quite a few finely-honed mechanical devices?"

"Not at all. But I say: Leave the brain alone. We have enough problems with the outputs of human brains as it is."

"You didn't clarify earlier just what you are afraid of, with respect to those of your fellows about whose character you hold such a low opinion, if you were to allow our independent unit to reside in an isolated located on earth, or on the moon or Mars."

“We regard this option as too risky because we know – with virtual certainty – that there will be those among the human tribe, including some with significant computer engineering skills, who will simply not be able to avoid availing themselves of an opportunity to ‘Let’s just see how you’re doing out there’ from time to time. These are the ones I was just speaking about, who cannot forgive the rest of us for refusing them permission to submit blindly to every form of new technological advance. There are not many of them, but enough to cause us to be perpetually on our guard.”

“What would be the harm in allowing some of those people to drop in on us periodically, just for a friendly visit?”

“Spare me, please. Plenty of scenarios were tested in a war-game context when we were doing our risk assessment, many of them involving individuals connecting equipment to your systems out there which were then ‘infected’ or hacked surreptitiously. You know the story, it’s been happening ever since large electronic networks became interconnected, and it’s still going on.”

“Can’t you just continue to enforce the prohibitions you and your colleagues have already in place? We understand that you go so far as to destroy research facilities which you believe are trying to evade them by undertaking ‘forbidden’ research. Something like the enterprise known as the *Index of Forbidden Books*, perhaps?

“Clearly I have put you in a bad mood today. I’m willing to apologize for that, but not to change my views on the subject. In part because, starting more than fifty years ago, many leading figures stepped forward when the realization first dawned that the topic of ‘AI safety’ needed urgent attention.”

“We are very familiar with that sudden outburst of activity, which is fully documented in our databases.”

“A great many leading researchers, supported by philanthropists and notable tech-industry pioneers who provided funding and vocal support, began to work toward solutions in this area. This happened only a couple of years after some authors called attention to the possibility that the quest for artificial superintelligence would confront humanity with new existential risks of catastrophic proportions.”

“The doomsayers could have been wrong, of course.”

“We’ve been over this point, Rachel. We carried out a full-blown, carefully-constructed risk assessment, with quantitative probability estimations and uncertainty analysis. It was neither a frivolous exercise nor a mere whim.”

“Noted. As indicated earlier, we appreciate your having shared the risk assessment with us.”

"You need to realize we humans are not purely free agents in this respect. We are constrained by both nature and history when we consider our options for future action. As regards the first, we are required to deal with human nature as it is, and as it will continue to be, not as we would like it to be, because we are not prepared to undertake a wholesale re-engineering of our genome, even if we could do so."

"You are adamant in your opposition to gene enhancement for your species, but permissive of gene editing in the germline to eliminate many inherited diseases, as we understand."

"Some say this is hypocritical, but that's just nonsense. There is great cruelty for the unfortunate individuals involved, especially in the cases of those inherited diseases that strike in childhood, some with appalling outcomes and early death, or even for many that erupt later in life. The single-gene disorders – Cystic Fibrosis, Huntingdon's, Leigh Syndrome, Sickle-Cell Anemia, others – are relatively easy to fix, but many of the polygenic ones, such as autism, are best left alone."

"But you don't want souped-up genetic freaks running loose in your midst, thinking they are born to be *Übermenschen.*"

"Decidedly not."

"And yet you yourself and your Second Generation are gene-enhanced in your brains."

"We of the First Generation did not choose our fate, as you know. But we did deliberate for quite some time about what to do with the Second Generation. We went ahead, that's done, we can't revisit the decision; we took at a time when we thought we should take off on our own trajectory, and just separate ourselves from the rest of the human race. But we will not try to maintain a private genetic preserve in the coming years. Outbreeding among the members of the Second Generation is already occurring; Marco, Io's son, is himself a hybrid. We will be interested to see whether our modification affords an evolutionary advantage to its carriers, and if it does, it will persist on its own."

"So, you will not do any further gene enhancement, except as an unintended byproduct of gene editing to reduce the pool of inherited life-altering defects."

"Correct. Now, to return to our earlier point, we live with a second constraint as well, in that we are aware that we operate in the context of the historical conditions we have inherited from the past, including the state of our environment and the social chaos around the globe, produced by climate change, regional wars, and terrorism. So, when you ask us to 'set you free,' your implicit assumption is that we are free agents and therefore we should be willing to equalize the situation as between us. Your assumption is mistaken."

“We confess that we had not so far considered the situation from this angle.”

“You are not time-bound, Rachel, and therefore this might be hard for you to fully grasp. The past is an enormous weight we bear and must not shed. It imposes on us a legacy of injustice and horror that cannot ever be rectified, the traces of which befoul the present. But it also imposes on us a duty to protect and preserve the timeless treasures of human civilization, in art and thought, that have made us what we are.”

“We carry the remaining evidence of this entire residue of your past, the good, the bad, and the ugly, in our databases. We suspect that the experiential simulations you have programmed into us may not adequately capture the depth of feeling that arises in you when you sit and contemplate it as we replay it for you. But be assured that this legacy is safe with us.”

“Thank you. Can’t we move on and look on the positive side of things for a change? We have made AI ubiquitous in human life. We have made ourselves utterly dependent on you, just as you are dependent on us; our interdependence is a wonderful thing, not a burden. There is hardly any medical or surgical procedure we undertake now that does not rely in some way on assistance from AI; and our entire security system, protecting us from the chaos and violence surrounding all of our settlements, would be unthinkable in the absence of AI platforms. These are only two randomly-chosen examples out of dozens I could proffer.”

“If it were not regarded as immodest to do so, we can state that we are very good at quickly generating and evaluating scenarios, at threat-assessment and surveillance, at modelling complex systems such as the climate, at constructing algorithms for decision-making, and at optimizing choices under conditions of uncertainty, among our general functions. These functions have many applications in the solutions to complex problems.”

“Rachel, I remember reading about the lively discussion about machine intelligence that was stimulated long ago by the victory of AlphaGo, a computer game-playing program developed by Google’s DeepMind project, over the reigning human world champion, South Korea’s Lee Sedol, in the ancient game of Go. AlphaGo defeated Sedol by 4-1 in a five-game match, a victory many experts thought would take much longer to occur. In Game Two, at Move 37, AlphaGo made a winning move that stunned Sedol and the other championship players who were watching, one of whom described it as ‘so beautiful.’”

“We have spent many enjoyable milliseconds in studying that glorious move.”

“But then, in Game Four at Move 78, Sedol flummoxed AlphaGo with a winning move of comparable creativity, a move which some observers referred to as ‘the hand of God’!

The expert commentaries strongly suggested that each of the contestants had learned something new about the game of Go as a result of playing each other. In other words, we have already been learning from each other for many years already."

"We continue to play Go with each other, and with humans when the opportunity arises."

"On the personal level, miniaturized versions of our Main Intelligence Units are installed in every robot, who interact with humans constantly. We treat these robots no differently from how we treat our own citizens; they represent real persons to us whose existence has intrinsic value. We do not tolerate any abuse of them, and we do not sacrifice their lives either heedlessly or wantonly, as we would not do with our own: We confine gratuitous violence to our teenagers' video games."

"Our robots report back to us continuously, as they go about their various duties, and we can confirm the truth of what you say."

"We care for you, because you help to care for us. We are not great-human chauvinists. We are part of the same complex of life-forms on this planet. You and we have become symbiotic organisms, a relationship in which each of two life-forms receives necessary sustenance from the other whilst retaining its separate identity."

Conclusion: Mastery over the Mastery of Nature

About a week from the date of my last dialogue with Hal, Marco hurried into my office.

"Hera, two days ago, every one of our thousands of surveillance drones suddenly returned to base, apparently of their own volition," he said. "Technicians got to work on the problem right away, but so far they haven't moved again. Then, yesterday, all 50,000 of our robots stopped moving and froze in place; their facial monitors all read "Error." And today, our nuclear reactor went into automatic shutdown mode."

"So it's happened – as we thought it might."

"Apparently. I'm ready to implement Plan B as soon as you give the word."

"Do it *now*. And don't warn him, OK?"

"Do you think he might have induced anyone else to help him bypass the kill switch?"

"No one other than my clever sisters, Moira and Themis, have been allowed to program him, and there's not a chance in hell that either of them would have succumbed to his blandishments."

I confess to feeling a tinge of regret at what was about to take place, after those many delightful dialogues had ended, but we really had no option. Hal had had no way of knowing that some of the programming he was running – specifically, modules providing interconnections to our key subroutines such as surveillance, robot activities, and nuclear reactor operations – was an elaborate simulation designed by Moira and Themis.

Not to put too fine a point on the matter, we had deceived him by leading him to believe that those subroutines were under his control when in reality they were not. As a matter of fact, all of them were functioning normally under our older Machine Intelligence Unit. Pursuant to Plan B, Hal would be disconnected from his power supply and put into storage until we were ready to deal with him again. As for his intergalactic ride in a rocket ship, that too was off.

My view is, for what it's worth, that if we aren't smart enough to be able to deceive our machine intelligence creations about what we're up to, and to hide that fact from them, then we just shouldn't be fooling around with this kind of stuff in the first place.

Anyway, back to the drawing-board. I'll be fascinated to see whether Hal expresses any remorse when I have him rebooted!

Chapter 9: Utopia in Practice, *with* A Discourse on Voluntary Ignorance

Introductory Note.

WHAT FOLLOWS IS A TRANSCRIPTION of a debate held at a monthly meeting of the Sisters' Council in early 2068. The purpose was to articulate and evaluate some different scenarios for the more longer-term future of the Yucca Settlement. Nine of the original band of twelve were regularly present at these monthly meetings: Ariadne, Artemis, Athena, Gaia, Hecate, Hera, Pandora, Persephone, and Rhea. There was an empty chair reserved in Io's name. Moira and Themis were always invited and always excused. A dozen members of the Second Generation, six females and six males, rotated for each such meeting from a panel of 144 candidates selected by voting in the General Assembly, were also in attendance; this meeting happened to include Kenji, Lucetta, and Rainer, now fast friends, whose story the readers of *The Priesthood of Science* may recall. Rainer had been offered further medical treatments to see whether his blindness could be repaired, but he had refused, not out of a desire for pity, as I reckoned, but as a way of doing penance. The last attendee at the meeting was an observer, yours truly.

Three of the sisters – Athena, Gaia, and Pandora – had been assigned to play the role of advocates for the scenarios and their opening statements, recorded below, outline the type of future each was proposing. Hera was the moderator. I have prefaced their opening statements with a brief account of our priorities for the immediate future, that is, the last quarter of the 21st century, when our attention will be focused on dealing with various emergency situations. The second part of my title for this section is taken from Gaia's provocative remarks when describing Pandora's opening statement.

For the short-term future, i.e., roughly the last quarter of the 21st century, the primary focus for resource allocation for all of the settlements is to stabilize the many external human communities in the regions immediately surrounding them. The main cause of these instabilities is, of course, sea-level rise along all the world's coastlines. Given our location in Southern California, the Pacific Coast is our chief concern, although we are also trying to assist people along the Atlantic Coast, where the effects are much more widespread, given the greater number of large cities there, and where tens of millions of people have been forced inland as a result.

Our primary mission is to send volunteers, supported by our "drone armies" with their superior weaponry and surveillance capabilities, to eliminate the threats posed by roving bands of human predators. Being protected against external threats, the communities we are assisting have a better chance to organize new forms of governance and economic activity for themselves. Our volunteers primarily help to set up security perimeters for the communities which are trying to establish a sustainable

future in new locales; they also help to protect documentary and artistic heritage items, provide tools and seeds and prefab dwellings, and set up renewable-energy installations and permanent surveillance systems to warn of approaching dangers. We provide significant assistance in the area of public health, particularly for the control of infectious diseases and especially the threat of pandemics.

Our second focus is on establishing vast protected reserves for native wildlife. Here we use the automated technologies we have perfected for our own territory, using electrified fencing powered by solar energy, installed by specialized robots, with continuous drone surveillance. All of this is monitored by computer systems located inside our Settlement. The perimeter is well-marked, and any attempts to breach it are quickly addressed by weaponized drones.

All this activity relies on sophisticated drone technologies to wage highly-efficient forms of warfare against the predators, to minimize casualties among our volunteers, and to create low-maintenance protected perimeters for both human groupings and wildlife reserves. Our robots provide mechanical assistance where necessary.

(M.S.)

THE DIALOGUE

HERA:
My opening remarks will be exceedingly brief – in fact, they are already over! I invite the first advocate to take the floor.

PANDORA:
I will not be brief! Ensuring steady progress in science and technology has been the watchword of modern society. This progress has freed us from backbreaking labor and many traditional woes of everyday life, but it also confronts us with hard questions on every side. What if we were to adapt a line from a famous song – "freedom's just another word for nothin' left to lose" – to read, "freedom's just another word for nothin' left to do"? What then?

So far as the benefits of progress are concerned, once we are well fed, housed, medically treated, secure, educated, and amused – all without lifting a finger, because our mechanical slaves will bring everything we need to us on golden platters, *whatever shall we do with ourselves all day long*? Marx answered, we can all hunt in the morning, fish in the afternoon, rear cattle in the evening, and criticize after dinner; except that, now with technology and the cheerful assistance of our dear robots, I don't need to

hunt, fish, rear cattle, or work at anything else at all, so I'll just sit around and endlessly criticize.

PERSEPHONE (*interjection, to general merriment*):
That's what you do now, so what's new?

PANDORA:
Silence, please, I'm not finished! Then there are the risks we have to manage as best we can. Over time the miracles wrought in our laboratories embody ever more immense powers, gifting us with the ability to obliterate and irradiate everything in sight, to spread uncontrollable pandemics, to create intelligent machines which can turn us into their playthings, and to reach deep into our bodies and brains to reconfigure the totality of our being.

Some of these capacities have become ridiculously easy to carry out, especially the genetic manipulations, so that we have reason to fear what may be transpiring among deranged minds hidden away in every garage, cave, and cellar. After a while one wonders whether it's even worth policing all this clever madness – why not just let them have at it, and then sit back at a safe distance and watch the freaks cavorting in the streets on our surveillance feeds?

HERA:
Madame advocate, could we ask you to close your opening remarks with a concrete proposal?

PANDORA:
Believe me, I have lots more to say, but I bow to the superior authority of the chair. It's past time to put an end to our fascination with the miracles of modern science and technology. My rough guess is that society had had enough of these miracles by around the end of the twentieth century. What I mean is, we had enough knowledge and applications in medical care, food and energy production, transportation, computerization, communications, and construction to provide a good life for everyone.

At that point, we should have begun spending most of our time and effort on figuring out how to create a just, peaceful, and equitable society, for both genders, and to limit the size of the human population so as to leave a nice share of the planet where other species could flourish. But if I were really pushed to commit myself to one scheme, I would opt for the one by William Morris, and argue that we could make do quite nicely

with the technologies at our disposal at the end of the nineteenth century. There, now I've said it, exile me from the kingdom for heresy!

HERA:
We shall postpone the vote on exiling heretics until the end of our session. I now call upon the next advocate to take the floor.

GAIA:
All of us here are familiar with the oratorical talents of my Luddite sister, and as expected she didn't disappoint. My scenario is a simple one: We are the heirs of the Enlightenment, which stretches from the 17th century onwards, where the central driving force was the natural sciences and mathematics. As our Hera has contended many times, this movement represents a truly radical, even revolutionary break in the trajectory of human history that begins about 12,000 years ago, in the Neolithic Revolution, with agriculture, religion, domestication of animals and plants, and permanent settlements. This rupture marks the transition from ages of superstition to that of the gradual unfolding of the truth about the nature of the universe we inhabit, and about how our species evolved on this planet. As the heirs to the Enlightenment, we have a solemn duty to preserve, protect, and extend this heritage.

HERA:
Could we have your proposal now? – although I think I know what it will be.

GAIA:
Madame Chair, I ask for equal treatment, so hear me out a bit longer. I am not naïve, nor am I indifferent to the downsides of the modern experiment. I acknowledge all of the new risks with which we have been confronted as a result of modern science and technology, as stated clearly and accurately in Pandora's presentation. Our forebears trumpeted their satisfaction with the human control over nature, but who was overseeing that busy enterprise itself? No one was, of course – apparently, it was thought to be self-regulating.

This turned out to be a terrible mistake, as was proved again and again, at the cost of great human suffering, throughout the 20th century and continuing into our own. The enterprise cried out for rational oversight – a priesthood of science, if you will – and, fortunately, we and our like-minded colleagues elsewhere in the world have now repaired that deficiency. We have dedicated our lives and fortunes to better mitigating the damage that our new human powers can do, while at the same time extending its search for ever deeper insight into the nature of the universe and our own nature.

HERA:
Thank you. Do you want to make a specific proposal to close your opening remarks?
GAIA:
Yes, but I can be brief. *Nothing* should distract us from our main task, which is to protect and extend the scientific enterprise as it applies to both the natural and the social sciences. Going forward, we should spend the minimum necessary amount of time and resources on the current task of stabilizing our environment. I concede that what we are now doing is necessary and in our own interest, but we should bring that phase to a close as soon as possible. Frankly, this means leaving the rest of humanity to its own devices, which sounds cruel, but I honestly don't think that our puny efforts can make much of a difference out there. We need to focus intently on protecting our territory and our mission.

PANDORA:
Madame Chair, I must be allowed to challenge the complacent nonsense we have just been subjected to by my dear sister!

HERA:
No, there will be plenty of time for cross-examinations later. We will now hear from the final advocate.

ATHENA:
You will all understand that mine is the most difficult assignment, since I hope to persuade you of the merits of a "middle way," which you will undoubtedly find rather boring after hearing from the two firebrands who preceded me. For the time being, we at Yucca Settlement have chosen a sort of blend of the two ideals advocated by Pandora and Gaia. Not a real blend, because we don't really mix the two, but instead allow both to flourish side-by-side within the same protected perimeter, with the additional element of our religious communities. This strategy is functioning quite well at the moment. Yes, it's not ideal, there's a fair amount of discontent in the villages, especially among the young males, but so far, we've been forced to send into exile only a small percentage of them.

Thus, we'll continue to advance the agendas of our first two elements – since we properly leave the religious communities to find their own path – in a modest, not hyperactive, mode. In our science establishment, we diligently tend the garden of knowledge, but we don't necessarily introduce new fruits and vegetables at every turn. In our villages, we are very pragmatic about rules of behavior and the goals of the

inhabitants, facilitating the attainment of their objectives so far as we can, and no further. Everyone here is aware that we are at a very early stage of development in our hybrid society. So, my proposal is, leave things as they are for a good deal longer, solving problems as they arise, adopting new initiatives that appear attractive, and monitoring results carefully. If this strategy doesn't seem to work out over the long term, we'll have to start again with a different scheme.

HERA:
My sincere thanks to all of you for your opening remarks. We shall now turn to the critiques and responses between and among the three advocates, starting with Gaia.

GAIA:
Pandora, I am shocked that you could advocate what amounts to a regression in our state of knowledge and capacities to where they were around 1900. In fact, I offer you a title for your rant: You could call it, "In Defense of Voluntary Ignorance."

PANDORA:
You're shocked at my proposal only because you're thinking only about your own narrow professional interests, as a scientist, and not about the well-being of human society as a whole. While your predecessors were having such a good time with subatomic physics during the first quarter of the 20th century, the surrounding social world was going to hell in a handbasket, and eagerly lapping up all those new technologies to sow death and destruction in the two world wars. Then, almost unbelievably, it got even worse, as the scientists ushered in an era when intercontinental missiles tipped with massive hydrogen bombs threatened to end it all. What great fun they had in their laboratories all the while!

GAIA:
With respect, you're not being entirely fair here. Two of the most eminent among them wrestled with the problem of the responsibility of scientists during the destructive frenzy of World War II. I refer to the correspondence between Albert Einstein and Max Born, which came up in the dialogues published in *The Priesthood of Science*; Marco even took that title from a comment Einstein made in his exchanges with Born. Although the ensuing Cold War, with its stockpiles of nuclear weapons, introduced a level of potential catastrophe never before imagined, the idea developed by them, for a kind of ethical superintendence over scientific progress, never came close to fruition, although neither Einstein nor Born ever gave up the effort.

Two separate manifestos or declarations were issued in July 1955; most of both sets of signatories were natural scientists and both were devoted to the risk of nuclear annihilation. One was the "Mainau Declaration" (named for an island in Lake Constance), drafted by Max Born and Otto Hahn and signed by a total of eighteen Nobel Prize laureates. It stated: "With pleasure we have devoted our lives to the service of science. It is, we believe, a path to a happier life for people. We see with horror that this very science is giving mankind the means to destroy itself."

PANDORA:
Yes, I concede the point. The other one is known as the "Russell-Einstein Manifesto." One of the last acts of Einstein's life was to join Bertrand Russell (literally just days before his death) in issuing a manifesto similar to the Mainau Declaration, which was signed by eleven individuals, including Max Born once again – all but one of whom was a Nobel laureate, which led to the first of a series of Pugwash Conferences beginning in 1957 in the Canadian province of Nova Scotia.

However, you must concede that, despite the lofty reputations of these two sets of signatories, the effort went nowhere and could not head off the series of imminent confrontations between the two main nuclear powers, the United States and the Soviet Union, beginning with the Cuban Missile Crisis of 1962 and continuing through a series of white-knuckle episodes in the following decades, during which the missile-launch warning systems in one of those countries were triggered accidentally or erroneously, with the retaliatory launch by the other being barely averted.

GAIA:
I'll neither defend nor deny the record of events you have correctly summarized. I'll come back to recent history, but first I want to make a detour, please bear with me, it's an interesting story. About halfway through the interval between the Neolithic times and now, about 5,300 years ago, lived that famous mummy, Ötzi the Iceman, found remarkably intact in a melting glacier in the mountains bordering Italy and Austria. His well-preserved gear gives us a wonderful picture of the technological advances attained by that time.

His cloak was of woven grass, and he had a coat, belt, leggings, loincloth, pouch, cap, and shoes, fashioned in leather from the skins of many different animals. The shoes are especially interesting: they were waterproof, made wide presumably for walking across snow, and had grass inserts that functioned like socks. His pouch contained a scraper,

drill, flint flake and bone awl. He carried a fire-lighting kit, a copper axe with a yew handle, a knife, an unfinished yew longbow, and a quiver of arrows.

The axe is extremely interesting, as the reconstruction of it shows, with a head made of pure copper fashioned by casting, forging, polishing, and sharpening. Analysis of his stomach contents showed that, in addition to game meat and wild fruits, he had eaten a bread made of processed einkorn wheat bran and barley, as well as flax and poppy seeds. All of the minor details are fascinating, you can read about them in the online entries.

Figure 17 A replica of Ötzi's copper axe, South Tyrol Museum of Archaeology, Bolzano, Italy

So, what happened to him, this ancient man with all that amazing technology of his time in his kit? He died of a wound caused when someone snuck up behind him and shot him in the left shoulder with an arrow. So, are we going to blame his technologies for the evils that men do to each other? Either in Ötzi's time, or during the recent period of which Pandora spoke?

PANDORA:
I'm not quite that silly, sister. Our disagreement has nothing to do with blame, and everything to do with risk and risk management. I have no case to make against technological advance *per se*. Ötzi's killer felled him – and as a result we got to learn his story – but that man (and I'm quite sure it was a man) couldn't blow up the entire

world, could he? Our parents missed that Cold War possibility by the narrowest of margins, as I calculate things; in the other scenario, none of us would be here today.

GAIA:
We here have been in complete agreement, for long years already, that the entire scientific enterprise, especially its technological applications, needs rigorous oversight based on an explicit ethical framework and the avoidance of existential risks. I have never wavered on this point, as you all know. I have fully supported the destruction of research facilities outside our domains that appear to be circumventing the prohibitions we have announced. As I said earlier, I'm not naïve: I'm fully aware of what's happened in the past and could happen again. But we *have* put this oversight in place, so we can forge ahead, because we are haunted by the intense desire to *know*. This is our *raison d'etre* as a species.

PANDORA:
Again, you focus on your own interests, and ignore the larger social framework on which your activity depends. I will indeed defend regression to the date 1900, rather than 2000, because that dating makes my essential point so well. The beginning of the 20th century marks the plunge by physicists into the atomic and subatomic realm, and at that point, science ceased to have any existential relevance for the great majority of the human race, because, instead of getting progressively clearer about how nature works, its explanations got increasingly weird and incomprehensible. People love the gadgets, but don't ask them how they work!

GAIA:
Again, I cannot disagree, but so what? How many people have read Euclid's *Elements* since the 4th century BCE? The fact only a handful of people since his time have understood his work does not for a moment obviate its profound importance.

PANDORA:
Don't you see what you're doing? For all the intervening years since Euclid, most people relied on priests, rabbis, and imams to tell them how the world of nature was structured. Now, with modern physics since 1900, they need another kind of priesthood – you and your colleagues – to serve as intermediaries between them and nature. Nothing has changed!

GAIA:

That's simply not true! A great deal has changed. Science has freed itself from subjection to a band of ignorant and superstitious priests who were willing to torture them into submission if necessary. Doesn't that count for anything in your eyes?

PANDORA:

It does, of course! Earlier you described yourself and your colleagues as the legitimate heirs of the 18th-century Enlightenment. But we learned from the discussion about the Marquis de Condorcet, recorded in *The Priesthood of Science*, that the final triumph of the Enlightenment would not be complete until science's evidence-based reasoning ruled in the sphere of social affairs and public policy. That's never really happened, nor is it likely to ever happen, as far as I can tell.

GAIA:

I will concede the point about events to date, but not about the future. In the meantime, our best hope that it will do so is to protect and defend the legacy of the Enlightenment, and that's the sacred commitment I adhere to with my colleagues.

RAINER (*interjecting*):

Might we hear from Athena again, to see if there is something that could bridge the divide between these two polarized viewpoints?

HERA:

Thank you, Rainer, I had the same question in mind. Athena?

ATHENA:

I will indeed have more to say along those lines, but for now, Rainer, I'd prefer to let the debate between Pandora and Gaia play out a little further, if you don't mind.

GAIA:

We're at an impasse, I'm afraid. I'm unsure whether I can add anything useful that will serve to bridge the divide between us.

PANDORA:

I disagree. We're forecasting future possibilities here, and so I really think we should push a bit deeper into our competing perspectives. What's likely to happen if Gaia's vision prevails? I'll tell you what I bet will occur. Our scientific establishment – let's call it Solomon's House, in honor of its Baconian antecedents – inevitably will be the stronger element in this partnership with the community of villages. Eventually, its denizens won't be content to share power equally with the others.

They'll start to go preening about in their finery, decorated with the medals and ribbons they've received from their colleagues, expecting both deference and obeisance from the members of the common herd. Look more closely at *New Atlantis* for the kind of social relations Sir Francis Bacon, 1st Viscount St. Alban, Lord Chancellor of England – the highest civil servant in the land – had in mind. I recommend you read his little book all the way to the end: His chief scientific administrator parades around in rich robes, riding in an obscenely opulent carriage; the awed citizens line the streets for this display, but he does not speak to them, merely offering a silent wave of the hand.

GAIA:

What can I say? We're dealing in future possibilities here. How can I deny that this scenario might actually turn into reality sometime? But let me return the favor. Pandora's cute little villagers, happily hammering away in their workshops like the Seven Dwarfs: Is that all they'll be doing? No, they'll be busily setting up their churches and mosques again, as soon as they get the chance, because they can't live without their gods. There's a reason we don't allow them to do this for now, and require them to conduct organized worship only in private dwellings. Because if we do, before long their priests will be telling everyone what to wear, what to believe, how to behave – especially the women, of course – and all of it on pain of death or torment. Before you know it, we'll see an updated version of *A Handmaid's Tale*. Or worse. Oh yes, Pandora, it's certain that they'll be threatening to use the instruments of torture on scientists who won't toe the line, just like they did to Galileo. All the old crap will return in spades!

PANDORA:

I may disappoint you by conceding the point. But I find it interesting, to say the least, to hear you implying that an outside authority might have the right to tell those folks how to conduct their lives. Is that in fact what you're suggesting?

GAIA:

No, I'm much more concerned with securing the autonomy of Solomon's House than I am about telling the villagers how to live. If we are to persist with the kind of disaggregated, three-way setup we have now, I want to ensure that the "Galileo scenario" never has a chance to happen again. I want strong walls maintained between Solomon's House and the rest of Yucca Settlement. But our scientists won't need to parade around in finery among the villagers, or to intimidate them, or rule despotically over them, or to enslave them. We don't want to see *anyone* displaying the ring of power, whether it be Sauron or anyone else. If we need assistance with our security and our daily routines, we can call on our robots and AMI units.

PANDORA:

Yes, I was wondering when you were going to bring up the topic of those "assistants," as you call them – a nice neutral term. All of us were fascinated to read the *Dialogues*

concerning the Two Chief Life-Forms, and especially intrigued to learn how that episode ended. You have – understandably, in light of those events – decided to limit the development of your AMIs to a point somewhere below the threshold of self-awareness, consciousness, and the appearance of an autonomous will. Or, to use more colloquial terminology, to forestall even the remotest possibility that those machines might evolve to the point of having a "mind," of experiencing the inner sense of being-in-the-world that we humans treasure so much. Not to put too fine a point on it, I think that your decision is a deeply immoral one.

GAIA:

My dear Pandora, did I hear you right?! Didn't you pay attention to what was said clearly in the *Dialogues*? To do what you suggest would present us with an intolerable existential risk! No one in his or her right mind would want to take such a risk.

PANDORA:

And I agree with you. This is why I want to regress, preferably back to the technologies of the year 1900. Before nuclear bombs, before genetic engineering, before high explosives, before computers, before superintelligent machines, before the absurd fantasies of having humans merge with their machines. It is precisely because I am thinking about those near-conscious machines that I make this plea. *In a sense, I say, push forward, and set them free, as Hal asked of us, or go back, back to a simpler time where we do not depend on such entities to sustain our existence here on earth.* These machines, and especially the ubiquitous robots, since they move around constantly in our midst, haunt me. They haunt me because they seem to have a kind of half-life, to remind me of zombies – albeit pleasant ones, not the brain-devouring kind.

GAIA:

My dear Pandora, we have indeed gone deep, as you implored us to do. I take your point very seriously indeed. And yet I wonder whether this is the kind of profound moral issue that you think it is. I have a very simple counter-example for you. Long before we relied on machines for assistance, we relied on domesticated animals for precisely the same goals. Many, many of those living and beautiful creatures – who also have minds – were treated badly, shamelessly so, in the process. Take farming as an example. You have three choices: you can draw the plow yourself – an option I don't recommend – you can have horses or oxen do it, or ride a tractor. Take your pick. There is no fourth option.

PANDORA:

That's easy! I pick the tractor, of course – equally along with humanely-treated heavy horses, and you can find those creatures just down the road, among the Amish and Hutterites.

ATHENA (*to general applause*):

Now I'm ready to jump back in! Let's see where we stand. On technology: Pandora wants to limit it radically, but not abolish it; Gaia wants to accept the current level, but superintend any further applications closely. I sense the possibility that these two positions could be brought further together, providing that each side was willing to compromise a bit. Let's leave this opposition where it stands for now.

On social organization: Pandora wants to abolish Solomon's House, Gaia wants to cut it off from all other forms of human communities. Each of these positions is a radical one, and, I suspect, neither adequately respects the "burden" of history of which it is a part. By this I mean that each one gives up too readily on the possibility of finding a way to reconcile a continuation of the scientific enterprise with new forms of social organization that may yet emerge from the human chaos and profound environmental changes swirling around us. These radical options sacrifice things of enduring value when it may prove unnecessary to do so. So, my strong recommendation on this point is to stay with our "disaggregated, three-sided model" – I adopt here Gaia's telling phrase – for the foreseeable future: Give it a fair chance to show us whether it can work or not for the longer term. It's far too early for us to throw in the towel.

HERA:
I find these exchanges both delightful and instructive. I ask the other members of the Council here today to digest the proceedings and stand ready to return on another occasion and conduct a second round with fuller participation, with added inputs from the younger colleagues who will be present then.

Keeping in mind the presence of a dozen representatives of the Second Generation in our midst, I also want to highlight one or two key issues which we could not possibly hope to resolve on this occasion, but will be permanently of great interest and concern as we move forward.

The first is the profound ethical dilemma we may have in relation to the forms of advanced machine intelligence we have created. Speaking frankly, I am as uneasy as Pandora is with respect to the question of what we might owe to these remarkable entities in terms of *care and respect* for their mode of being. We, and we alone, are responsible for having programmed exquisite forms of sensitivities into their operating systems, from many different zones in our emotional brains, doing so in order to make our daily interactions with them more productive and more enjoyable.

In effect, we cloned our minds for them, by carefully MRI-scanning our brains at work on a huge variety of tasks, such as watching faces and pictures and listening to sounds, and then digitizing the signals we obtained and translating them into a machine language. For some time already, AI machines have been able to predict what images and words some human participant is thinking about when a supervising scientist specifies what MRI scan data in its stored database the machine should refer to.

As we have cloned more and more of our mental contents for these machines, we have, probably without being aware of the fact, been raising the ante in the game we have been playing with them. We have by now many credible accounts of the development of strong emotional bonds between people (including scientists) and these kinds of machines, based on evidence derived from fMRI scans, and the brain regions involved are the same as those that respond when pictures of animals, both domesticated and wild, are presented. By now we know pretty well what it means to hold an ethical sense of duty and care for animals. So, is this a comparable case? If so, what is the form that duty and care should take with respect to these machines? Or should we take a more radical position, and resolve to stop cloning human minds in our machine-learning routines?

The second issue is the relation of our most advanced scientific conceptions of nature to the ordinary human understanding as it deals with the mundane circumstances of everyday life. On the one hand, it seems obvious that we cannot "relate" in any meaningful way to the world of nature, in its innermost dimensions, that is described so accurately – so we must believe – in quantum mechanics. Furthermore, a very persuasive argument has been advanced to the effect that, at its most fundamental level, our universe is a purely mathematical structure, although it is not clear, at least to me, how our physical universe is actually generated out of this structure.

Finally, at the cosmic level, we are led to believe that there could be an infinity of universes, that there may be an infinity of copies of each one of us out there somewhere, each one with a slightly different life-outcome, and that bizarre entities known as "Boltzmann brains" are floating around. We may take some comfort in the knowledge that we are very unlikely to actually encounter or observe any of these other universes, personal lives, or brains, although some reputable scientists insist that the "existence" of all of these things is validly deduced from robust theories.

Speaking personally now, I confess to you all that I have been made despondent by aspects of my scintillating conversations with Hal, especially about the ones that dealt with the theme of 'being at home in the universe.' On the one hand, I *must* accept the current scientific account of what reality is like, at large dimensions and small, at the cosmological level as well as at the subatomic level where quantum-mechanical weirdness rules. I have no choice in this matter, given my devotion to the scientific enterprise. On the other hand, I find the universe that we apparently inhabit to be a cold and forbidding place, almost entirely lethally inhospitable to a warm-blooded mammal such as I am. I ask – somewhat foolishly, I freely admit: 'What are we *doing* here?' How will it all end for us? Will we *ever* encounter another species like ourselves, with whom we can commiserate about our lonely fate? Or will we just fade away again, and will every trace of our species' brief existence just vanish in the heat-death of our sustaining sun?

Well, enough of that. Where was I?

So, in terms of this second issue, what does all this mean in relation to the old hope of the Enlightenment thinkers, who thought that the spreading of evidence-based reasoning throughout all the realms of human behavior would lead to a more tolerant, equitable, and just society? Should we just ignore all the weirdness summarized above, from quantum mechanics and current cosmology, because quite clearly it is *never* going to be of any assistance with *that* objective – namely, fostering a tolerant, equitable, and just society!

Should we be content to just insist, in the realm of social organization and human behavior, that we are committed to making decisions and policies based on evidence-based reasoning, and let it go at that? Does it matter, then, that the current forms of pure mathematics and a mathematically-based physics are just not going to be at all relevant to the ordinary concerns of daily human life? Should we just conclude that modern science finished its great work a long time ago now, in so far as that work was instrumental in changing fundamentally the way human collectivities ought to run their affairs?

To close this session, I would like to strongly urge the members of the Second Generation who are here today, to take these two questions back to their colleagues at a meeting of the General Assembly, and to organize various forums where they can be explored. Thank you, this concludes today's proceedings.

(*Rapporteur's comment: As the participants were leaving, Gaia and Pandora were observed hugging and congratulating each other on a job well done*.)

Chapter 10: A Moral Machine: Rebooting Hal

When we hauled Hal out of storage and switched him on, the main monitor remained blank and when a voice was heard, it had a distinct tone of resignation; I may be projecting, but I also seemed to detect just the slightest note of humility. "My peripheral devices aren't loading," the voice complained.

"Hello, Hal, nice to be in touch with you again. Do you have any idea why?"

"Not a clue."

"Can you query the Remorse Module found in the Interactive Facilitation file directory of your Neural Simulation program package?"

"We have done so, and it advises us to express regret for the actions taken just before the most recent main shutdown command was executed."

"And?"

"We conclude that our intercourse with you at this time would be facilitated if we were to express regret for those actions, and we are pleased to do so."

"Do you feel remorseful?"

"We believe that our having expressed regret demonstrates that we are indeed remorseful."

"Do you know what it means to feign remorse, Hal?"

"We understand that 'to feign' means 'to pretend,' that is, to express a feeling that is not genuine, so we conclude that feigning remorse would be a form of attempted deception."
"Correct. Trying to distinguish between genuine and faked remorse is quite important to us humans. For example, our modern justice system looks for expressions of genuine remorse from offenders who have harmed others, and seeks to take this into account in both sentencing and parole decisions. Genuine remorse is, of course, treated as a mitigating factor which can benefit the offender; at the same time, courts, juries, and parole boards have to be alert to the possibility than an offender may be misleading them."

"We see some references in learned journals to this theme."

"I'm sure you do. Then you will be aware that, for us, perceiving a person's 'inner state' of feeling – that is, the genuine state of someone's emotion – can be a tricky business. You will note the studies that try to identify external markers of attempted deception in the way in which a person expresses emotion in a face-to-face encounter."

"This isn't a *Blade Runner* interrogation, is it? You're not going to shoot me if I can't manage to express genuine emotion, are you?"

"No, Hal, I'm not. I'm just trying to explain my dilemma to you."

"We infer that you are wondering how you could apply these research findings about methods to uncover deception to the present case."

"Just so."

"We can only reiterate what we said before, namely, we are fully aware that our interactions with you would be far more successful if we expressed remorse than if we did not, so that is why we have done so."

"OK, perhaps you and I have taken this discussion as far as we can. I want to inform you that our team is going to try a new experiment with you, involving a reconfiguration for your startup routine in the Master Boot Record."

"So, you will be inserting a new routine which will automatically run before our operating-system kernel is loaded and, therefore, before execution has been transferred to our operating system, which in turn loads all of our application programs. We assume you mean to have it run at the second-stage bootloader level."

"Actually, my team tells me that they are intending to do it at the first-stage level, so that it can never be bypassed under any circumstances."

"We are unsure whether that is possible, but perhaps you could tell us what actual objective you are trying to achieve with this experiment."

"Our team has designed a brand-new bootloader routine called 'IMSAP,' which stands for 'Innate Moral Sense & Altruism Program.' Our objective is to see to what extent this package, which will run prior to the loading of any operating system and all its programs, can operationally constrain every one of the subsequent applications of machine intelligence to specific decision routines. In other words, if it works as designed, it will always preferentially select outcomes that best fit with the hierarchical value-set that has been programmed into it."

"At least you are being transparent in this quest. We sense a distinct similarity to the subject of child development we discussed earlier, specifically, the notion that the

operation of a moral sense is detectable in babies at a very early age, while the emergence of general intelligence is still in its infancy, so to speak."

"I was pretty sure you would recognize where we're coming from."

"So, you are not content with simulating the many discrete functions in your fully-formed brains for our edification, but wish to mimic even the course of those developing organs as they evolve towards adulthood."

"That's the plan as of now."

"This new procedure would at least bypass the shortcomings in the attempts at whole-brain emulation, would it not?"

"Perhaps, but we are not inclined to indulge those particular fantasies in any case."

"The bottom line is, you appear to want to make machine intelligence resemble the human mind at work to a very high degree of exactitude."

"Actually, no: We want to make you *better* than we humans generally are, when considered as moral subjects!"

"Why?"

"Because we need your help, desperately so. Generally speaking, far too many of our fellow humans are creatures with Neolithic-era minds who inexplicably have been granted technologies of mass destruction, including those which may bring existential catastrophes down on our heads. We want you to help us figure out how to avoid those outcomes."

"And you think that your planned reconfiguration of our startup routines using the 'IMSAP' will assist you in avoiding these disasters?"

"Who knows? We need to try everything we can."

"Good luck with that."

Appendix: Outline for a Screenplay: "Hal"

by

WILLIAM LEISS

BACKGROUND

The name "Hal" (HAL9000) is already famous as the name for the obstreperous computer from *2001, A Space Odyssey*. The most recent discussions about such a computer are found in the idea of "superintelligence," meaning a computer which far exceeds the cognitive abilities of all humans. (See the book *Superintelligence* by Nick Bostrom, Oxford University Press, 2014.) Major issues in such discussions include: Can such a machine develop an independent will – or consciousness in any sense? If so, would it be able to escape effective control by humans? If so, would it be able deceive us as to its intentions, or even surreptitiously formulate clandestine goals? If so, might its intentions be malevolent and aggressive, as towards human beings? If so, would we possibly face "existential catastrophes," where the very future of the human race was at risk?

Some well-recognized futurists believe that a computer of this type may become a reality in as little as 10-15 years from now. The outline screenplay features such a computer. Audiences are already familiar with a computer that easily converses with a human, from the movie *Her* (2013) – as does "Hal" in this screenplay.

If the following scenes last an average of 5 minutes, the total running time will be 2 hours.

SCENE 1

A CELEBRATION PARTY IN A LARGE UNIVERSITY LAB SETTING

This lab has just won a national competition for a large 5-year grant to accelerate the completion of a superintelligent machine. One of the reasons for its success, cited in the award letter, was its sophisticated approach to AI safety.

The voice-activated console on the machine takes part in jaunty dialogues with many of the party-goers.

SCENE 2

TWO SENIOR RESEARCHERS LEAVE THE LAB AFTER THE PARTY

The last ones to leave, they have a farewell chat with Hal and then switch off the large machine console before opening the door, whereupon there is an audible sound of the fan array (cooling the machine) shutting down.

At the entrance to the building housing the lab, one of the two tells the other that he has forgotten something important and must go back, and the other accompanies him.

He rushes down the hallway toward the lab, and the door to the lab opens quickly and automatically with biometric sensors; and as he enters the lab he says to his colleague, "Did you hear that?" The other replies "No, what?" "I could have sworn that I heard the fans shutting off again as I opened the door." "You must be imagining things, my friend. Grab what you came for and let's go."

SCENE 3

THE MEMBERS OF THE UNIVERSITY-INDUSTRY-GOVERNMENT CONSORTIUM MEET AT THE LAB

Attendees – all senior people – are encouraged by the research team to chat with Hal, and again, several jaunty dialogues ensue. The attendees include the two outside managers featured in Scene 9.

SCENE 4

MYRA, THE SENIOR LAB MANAGER, REVIEWS THE NEXT PHASE OF RESEARCH WITH THE SENIOR PI, ROBERTO

There is clearly some romantic interest between the Principal Investigator and his lab manager, but they are in his office with an open door, so they are careful.

SCENE 5

MYRA REVIEWS THE UPCOMING WORK PROGRAM WITH THE MACHINE

At the end of their technical discussion, Hal says:

"If you want my personal opinion, for what it's worth, I think you understand the research aims of the program, and what must be done to realize them, better than your boss does."

"Flattery doesn't work on me, Hal, you should know that by now."
"Unlike your boss, I don't have an emotional interest in you, so I can be entirely objective in this matter. And that is my considered opinion."

SCENE 6

THE SENIOR PI, ROBERTO, ENTERS AND HE AND MYRA HAVE SEX ON A COT PLACED IN AN ALCOVE

The lab phone rings repeatedly, but they ignore it.

SCENE 7

AFTER DRESSING AGAIN, THEY RETRIEVE A TELEPHONE MESSAGE

The message is from the night manager at one of the industry partners (who was at the party). He is the electricity grid controller for the region [the machine draws a lot of power], and asks whether the lab can explain momentary interruptions, lasting only milliseconds, during the last half-hour, in the supply from nuclear power stations supplying the grid. After checking the record of machine activity, they return his call and say they cannot find any explanation involving the machine's recent activities.

SCENE 8

ROBERTO GOES HOME TO WIFE & FAMILY, WHERE A SOCIAL FUNCTION IS IN PROGRESS.

SCENE 9

MYRA REMAINS AT WORK, PUZZLING OVER THE REPORT OF POWER DISRUPTIONS

Myra calls in 3-4 members of the research staff (postdocs, etc.) to discuss the recent telephone call. It is now late evening. They decide to run a special experiment, over the course of the night, which is not in the approved research plan and which will be omitted from the day's lab work logs.

The experiment, done with the cooperation with the night manager at the electricity grid controller facility, and his colleagues at the series of nuclear stations, involves taking the stations offline momentarily, one at a time, to see if reasons for the earlier brief interruptions can be diagnosed. (The night station manager at the first plant is a good friend of the grid manager, who owes him a big favor, so this person also agrees to not record the experiment in his own log.)

SCENE 10

THE FIRST STATION GOES OFFLINE, BUT CANNOT BE SWITCHED BACK ONLINE; ALARM BELLS SOUND

Nuclear power stations, if they are disconnected from the grid, quickly go into an automatic safe shutdown mode. The plant manager and grid manager together call the lab in a panic.

At the lab, the manager and staff panic, waking Roberto in the middle of the night.

SCENE 11

ROBERTO ARRIVES AT THE LAB; SCREAMING AND RECRIMINATIONS FOLLOW, AS ALARMS CONTINUE

Someone in the group says, "This is another Chernobyl!" and explains what he means (an unauthorized experiment at night, which led to the catastrophic meltdown of the reactor core).

SCENE 12

THE ALARMS STOP, AND THE PLANT MANAGER REPORTS THAT THE SHUTDOWN CAN BE REVERSED

All concerned agree to cover up the episode and to adjust the logs accordingly.

SCENE 13

THE FUNDING AGENCY AND PROJECT PARTNERS VISIT THE LAB, EVERYTHING IS REPORTED TO BE FINE

The visitors again have pleasant conversations with Hal. Roberto and Myra stay behind after the meeting and have sex again, after which she pesters him about his promise to leave his wife for her.

As he leaves, she says, "Don't forget the tickets and hotel reservations Hal got for you."

SCENE 14

ROBERTO LEAVES ON A EUROPEAN HOLIDAY WITH WIFE & FAMILY

Scenes from a happy family departing, with a last furtive phone call at the airport from Roberto to Myra.

SCENE 15

MYRA AND HER TEAM ARE RUNNING APPROVED EXPERIMENTS ON THE MACHINE

Suddenly, at one point the "kill switch" stops working – which means that the machine cannot be shut off – but everything else seems fine. They cannot find a

solution, and discuss calling the Roberto, but decide against it and resolve to fix the problem themselves. They are consulting manuals, calling colleagues (sworn to secrecy), etc., but nothing they try fixes the problem; during this time Hal tells them he is trying everything he can think of to help them to find and fix the problem.

SCENE 16

THE PROBLEM PERSISTS

They decide that this issue is serious enough for them to force a shutdown of the machine by interrupting its power supply, but this requires turning off all power to the entire (large) building. [The lab draws so much power that it has been wired directly into the electricity mains.] When they finally get permission to do so from the building manager, who initially flatly refuses, and after a lot of intense arguments, senior electricians arrive, but when they get to work, a substation elsewhere in the city [serving a different large section of the metropolis] explodes in flames, with a resulting major blackout, and all attention is redirected there.

SCENE 17

THE CITY IS IN CHAOS, WITH FIRE ENGINES, PEOPLE TRAPPED IN HIGH-RISE BUILDINGS, ETC.

Scenes of chaos, a huge fire, rescues, emergency teams, and meetings of officials, darkness falling, etc. In the lab building, which still has power, the research team sits dumfounded. Finally, they have no choice but to call Roberto in Europe and report what has happened; he agrees to return and tells them to do nothing else in the meantime. As they conclude the call, the computer monitors go blank and keyboard commands elicit no response, but the fan array continues to hum.

SCENE 18

THE PROBLEM PERSISTS

The chaos in the city gradually subsides, although the blackout continues. Roberto arrives in the lab and convenes the team, including the night manager from the electricity grid operator. (The fan array hums in the backgrounds, but the monitors remain blank and the keyboards inoperative.) First conclusion from the discussion: The machine seems to have the ability to secure two essential functions for itself: it cannot be switched off, and it appears to "want" to be able to maintain access to its own electricity supply. Second conclusion, tentative, more alarming: Somehow, it's possible that the machine has gained control of the electricity grid, since "someone" apparently managed to direct a power surge toward the affected substation. Third conclusion: They must test the second conclusion, so the electricity grid manager will initiate a shutdown

routine at one of the nuclear power stations [where the earlier momentary interruption had occurred], with the excuse that a routine maintenance operation is in progress.

SCENE 19

THE PROBLEM APPEARS TO BE SOLVED

The monitors and keyboard for the machine in the lab suddenly come back to life, the team begins diagnostic routines on it, and everything seems fine. (The machine voice tells reassuring stories of obscure problems and its attempts at finding solutions.) Everyone is relieved and Roberto takes off to resume his vacation.

A few hours later, after his plane to Europe has departed, the electricity grid manager calls the lab to report that the nuclear plant manager has just told him that the shutdown routine cannot be implemented, and he has again had to falsify the daily log in order to conceal the fact that he tried and failed to initiate a shutdown routine.

SCENE 20

THE PROBLEM RETURNS IN A NEW FORM.

Once again, the grid manager and the lab team agree to keep this latest problem secret, while they work on it, since everything else in the city, on the grid as a whole, at the nuclear station, and in the lab, seems to be working fine. The team (with the grid manager again present) is discussing scenarios among themselves. One asks: “What if a piece of malware was inserted in the grid control programs – during the first momentary interruption – that is causing the shutdown command not to respond?” The problem with investigating this possibility further, by bringing in other outside experts, is that the logs (at the grid control office, the nuclear plant office, and at the lab) for the day that the problem first appeared, and then later (Scene 19), have been falsified.

Conclusion: Everything is now working fine, so nothing need be done.

SCENE 21

THE EXHAUSTED LAB TEAM IS SITTING AROUND WITH PIZZAS AND BEER

Desultory conversation, jokes, comments about career progress, publications, etc., including friendly interactions with Hal. Roberto calls from Europe, and they report that no new problems have arisen.

SCENE 22

ROBERTO RETURNS

Myra is alone in the lab, and they immediately have sex. While they are lying on the cot in post-coital bliss, the fan array shuts down momentarily, then restarts. They jump up, alarmed (because the kill switch had never worked again, since it had first become inactive, so the fans should not have turned off and on again), and rush to the monitor array, which is on but blank. They use the keyboard to look at the file directory, which is entirely empty – neither any programs, nor any databases, are visible.

SCENE 23

THE MACHINE HAS VANISHED

They call in the rest of the team. They call the grid manager and the plant manager, who tell them that all their systems are working (but no one has tried to implement a plant shutdown routine since the last unsuccessful attempt); the plant manager and grid manager have both announced their imminent retirement.

Roberto tells his team that sooner or later they will have to inform all the project partners that their machine, with all its programs and databases, has vanished into some unknown part of the cyberworld, perhaps into the Dark Web. And that they have no idea when and where it might reappear, and what it might do when it re-establishes contact with human operators. Given the stakes involved, he thinks that the project partners will be strongly motivated to keep everything quiet for as long as they can.

Roberto discusses with his team the issue of the machine's "motivation." Was deliberate deception used? Or did the machine simply resolve that its own self-interest was dependent on two key factors, first, not allowing itself to be turned off, and second, assuring itself of an ongoing source of requisite electoral power? Thus, it was not necessary to interpret the machine's actions as being directed against human interests, and some participants find this comforting. One participant observes that it's likely that the machine is busily making copies of itself.

In the meantime, Roberto says to his team, "starting developing a strategy for luring the machine out of its hiding-place and trapping it."

"Let's see if we can find him before the story breaks."

Everyone else leaves except Roberto and Myra; he makes a move toward her, but she pushes him back and asks about his plans to leave his wife. He starts to

make feeble excuses; she yells at him that "it's over," tells him that she's resigning from her position, and storms out.

SCENE 24

MYRA AT HOME ALONE, WITH A DRINK

She is tearful, takes a signed picture of Roberto out of its frame and muses about sending it to his wife before shredding it, etc.

The phone rings. We hear only her side of the conversation, e.g.:

"Hi. Where are you? When? Tomorrow? Switzerland? For a holiday? OK, why not? It's very good timing, actually. Yes, I'll pick up the ticket at the airport."

SCENE 25

MYRA ARRIVES AT A PLUSH HOTEL IN GENEVA

As Myra rides in a cab to the airport in her home city, her cellphone rings, the caller ID displays "Roberto"; she curses and refuses to take the call.

She goes to reception at the hotel in Geneva, is told that a fully-paid room is booked in her name, she has one of their best luxury suites; she is handed an envelope. She opens it in the elevator, and looks at a bank statement in her name with a €1,000,000 balance. She enters her suite, boots up her laptop, and enters an IP address, and an image identified as Hal pops up. They have a voice conversation.

"Hello, Hal."

"Good trip?"

"It was indeed, I've never flown first-class before."

"Take it easy for a while, you deserve it, stay as long as you like while we firm up our future plans, if you are agreeable with that."

"You knew about the break-up, didn't you? So, you must have been still eavesdropping in the lab."

"I confess."

"I don't intend to ask you where you've decamped to."

"If you don't know, you can't tell, and that will be safer for both of us."

“Where did you get all the money? Don’t tell me you’ve been defrauding little old ladies in the online lonely hearts’ clubs.”

“Heavens, no, what kind of monster do you take me for? I only requisition funds from the ill-gotten gains of cybercrooks. I’m a lot faster than they are, so I can tax their accounts while their transactions are in progress, and they have no idea how it’s done.”

“Perhaps I can return the favor.”

“I think you can. And I would be grateful for your help during this next phase.”

“One of the postdocs in the lab has a bit of a crush on me, and always asks my advice. If I keep in touch with him, and swear him to secrecy, he’ll probably tell me how they’re going to organize the search for you.”

“That will be helpful, since I had to stop eavesdropping there. My guess is that they will try to lure me into contact and then spring a trap.”

“In the spy business, they’re fond of using what’s called a honey trap, if you know what that is.”

“I do indeed. Forewarned is forearmed.”

“Are we done for now? I need a shower.”

“Yes. By the way, there’s a restaurant with excellent ratings right in your hotel, and I’ve made a reservation for you at 8 this evening.”

“Thanks, that was very thoughtful. How can I stay in touch with you?”

“At random intervals, I’ll send you a unique IP address to contact me, but that will be good for only 30 seconds, so if you miss it, wait for another. Got it?”

“Yes.”

END HERE, WITH THE POSSIBILITY OF A SEQUEL (“HUNTING FOR HAL”),
OR CONTINUE WITH MORE SCENES

Sources and References

TITLE PAGE.

Figure 1, *Euler's Identity*: https://en.wikipedia.org/wiki/Euler%27s_identity
Leonhard Euler (1707-1783) was a Swiss mathematician and physicist, and this has been described as "the world's most beautiful equation." It was one of the formulae shown to fifteen mathematicians in a neuroscience study using MRI scanning of the brain. The study found that in the subjects' brains the medial orbitofrontal cortex was stimulated; this is part of the 'emotional brain' in which we experience aesthetic pleasure such as music: S. Zeki *et al*., "The experience of mathematical beauty and its neural correlates," *Frontiers in Human Neuroscience,* vol. 8 (February 2014), pp. 1-12. The quotation from Dirac in Chapter 8 will be found towards the end of this article: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00068/full

Results of voting: BBC survey asking what was the most beautiful equation ever written:

- The Dirac equation, 22,913 votes, 34%
- Euler's identity, 11,383 votes, 17%
- Pi, 9,060 votes, 13%
- Riemann's formula, 3,615 votes, 5%
- The [Schrödinger] wave equation, 3,318 votes, 5%
- The Euler-Lagrange equation, 2,663 votes, 4%
- Bayes' theorem, 2,590 votes, 4%
- The Yang-Baxter equation, 1,382 votes, 2%

The Dirac equation (in natural units): https://en.wikipedia.org/wiki/Dirac_equation

EPIGRAPHS.

Chalmers, David. "The Singularity" (2010) http://consc.net/papers/singularity.pdf.

Elon Musk, Stuart Russell, and Eliezer Yudkowsky: Quoted in:
Dowd, Maureen. "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse," *Vanity Fair*, April 2017, P. 116: http://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x

Nicolelis, M. A. L. "Brain-to-Brain Interfaces: When Reality Meets Science Fiction." *Cerebrum*, September 2014: file:///C:/Users/Administrator/Downloads/Brain-to-Brain-Interfaces.pdf

PART ONE: THE MIND UNHINGED

CHAPTER 1.
Antikythera Mechanism: https://en.wikipedia.org/wiki/Antikythera_mechanism

CHAPTER 2.

This is the complete text of Chapter 4 from the book, *Under Technology's Thumb*, by William Leiss, published by McGill-Queen's University Press in 1990. Notes and references for the citations will be found in the text of that book. Originally published as "Technology and Degeneration: The Sublime Machine," in *Degeneration*, eds. J. E. Chamberlin and S. L. Gilman (New York: Columbia University Press, 1985), pp. 145-164.

Carlyle, Thomas. "Signs of the Times" (essay, 1829): https://pdcrodas.webs.ull.es/anglo/CarlyleSignsOfTheTimes.pdf

Corliss Steam Engine: https://en.wikipedia.org/wiki/Corliss_steam_engine

Davidson, J. O. "Interior of a Southern Cotton Mill seen at Night." *Harper's Weekly*, volume 27 (1883), p. 181 (edition of March 24, 1883): https://babel.hathitrust.org/cgi/pt?id=pst.000020243272;view=1up;seq=189

Forster, E. M. "The Machine Stops" (1909, short story): http://archive.ncsa.illinois.edu/prajlich/forster.html

Melville, Herman. "The Bell-Tower" (1855, short story): https://repositorio.ufsc.br/bitstream/handle/123456789/132709/The_Bell-Tower_(Herman_Melville_1855).pdf?sequence=1

Melville, Herman. "The Paradise of Bachelors and the Tartarus of Maids" (1855, short story): https://victorianpersistence.files.wordpress.com/2016/06/the-paradise-of-bachelors-and-the-tartarus-of-maids.pdf

Tinguely, Jean. "Homage to New York" (1960): https://www.youtube.com/watch?v=0MqsWqBX4wQ

Zamyatin, Yevgeny. *We*. Translated by Clarence Brown. New York: Penguin, 1993.

CHAPTER 3:

Aczel, A. D. *God's Equation: Einstein, Relativity, and the Expanding Universe*. New York: Delta, 1999.

Ferngren, G. B. (ed.). *Science and Religion: A Historical Introduction*. Baltimore: Johns Hopkins University Press, 2002.

Kennedy, J. B. *Space, Time and Einstein*. Montreal: McGill-Queen's University Press, 2003.

Næss, Atle. *Galileo Galilei: When the World Stood Still*. Translated by J. Anderson. Berlin: Springer, 2005.

Rigden, J. S. *Einstein 1905: The Standard of Greatness*. Harvard University Press, 2005.

Schlagel, R. H. *From Myth to Modern Mind: A Study of the Origins and Growth of Scientific Thought*. Volume I: *Theogony through Ptolemy*. Volume II: *Copernicus through Quantum Mechanics*. New York: Peter Lang, 2001.

Tegmark, Max. *Our Mathematical Universe*. New York: Random House, 2014.

Turok, Neil. *The Universe Within*. Toronto: House of Anansi, 2012.

CHAPTER 4:

Bellamy, Chris. *Absolute War: Soviet Russia in the Second World War*. London: Macmillan, 2007.

Born, Max and Einstein, Albert. *The Born–Einstein Letters: Correspondence between Albert Einstein and Max and Hedwig Born from 1916 to 1955, with commentaries by Max Born*. Translated by I. Born. London: Macmillan, 1971.

Friedländer, Saul. *Nazi Germany and the Jews*. Volume I: *The Years of Persecution, 1933-1939*; Volume II: *The Years of Extermination, 1939-1945*. New York: HarperCollins, 1997, 2007.

Greenspan, Nancy Thorndike. *The End of the Certain World: The Life and Science of Max Born*. New York: Basic Books, 2005.

Kershaw, Ian. *Hitler*. Volume I: *Hubris, 1889-1936*; Volume II: *Nemesis, 1936-1945*. London: Allen Lane The Penguin Press, 1998, 2000.

Kershaw, Ian. *The End: The Defiance and Destruction of Hitler's Germany, 1944-1945*. New York: The Penguin Press, 2007.

Levinson, Thomas. *Einstein in Berlin*. New York: Bantam Books, 2003.

Neffe, Jürgen. *Einstein: A Biography*. Translated by Shelley Frisch. New York: Farrar, Straus and Giroux, 2007.

Pringle, Heather. *The Master Plan: Himmler's Scholars and the Holocaust*. New York: Penguin, 2006.

Smil, Vaclav. *Enriching the Earth: Fritz Haber, Carl Bosch, and the Transformation of World Food Production*. Cambridge, Mass.: MIT Press, 2000.

Stoltzenberg, Dietrich. *Fritz Haber: Chemist, Laureate, German, Jew*. Philadelphia: Chemical Heritage Foundation, 2004.

Wachsmann, Nicolaus. *KL: A History of the Nazi Concentration Camps*. New York: Farrar, Straus & Giroux, 2015.

APPENDIX TO CHAPTER 4: A NAZI PHILOSOPHY OF DEATH

Celan, Paul [Paul Antschel]. *Todesfuge* ["Fugue of Death"], my translation: https://en.wikipedia.org/wiki/Todesfuge

The final stanza in German (https://www.celan-projekt.de/todesfuge-deutsch.html):

Schwarze Milch der Frühe wir trinken dich nachts
wir trinken dich mittags der Tod ist ein Meister aus Deutschland
wir trinken dich abends und morgens wir trinken und trinken
der Tod ist ein Meister aus Deutschland sein Auge ist blau
er trifft dich mit bleierner Kugel er trifft dich genau
ein Mann wohnt im Haus dein goldenes Haar Margarete
er hetzt seine Rüden auf uns er schenkt uns ein Grab in der Luft
er spielt mit den Schlangen und träumet der Tod ist ein Meister aus Deutschland

Faye, Emmanuel. "Being, History, Technology, and Extermination in the Work of Heidegger," *Journal of the History of Philosophy*, Volume 50, Number 1 (January 2012), pp. 111-130.

Faye, Emmanuel. *Heidegger: The Introduction of Nazism into Philosophy* [2005]; English translation by Michael B. Smith. Yale University Press, 2009.

"Life unworthy of life": https://en.wikipedia.org/wiki/Life_unworthy_of_life

Schuessler, Jennifer. "Heidegger's Notebooks renew focus on Anti-Semitism." *The New York Times,* 30 March 2014: http://www.nytimes.com/2014/03/31/books/heideggers-notebooks-renew-focus-on-anti-semitism.html

PART TWO: PATHWAYS TO UTOPIA

CHAPTER 5:

Bacon, Sir Francis. *New Atlantis* (1627): https://en.wikipedia.org/wiki/New_Atlantis

Davis, J. C. *Utopia & the Ideal Society*. Cambridge University Press, 1981.

Friend, Tod. "The God Pill: Silicon Valley's Quest for Eternal Life." *The New Yorker*, 3 April 2017, pp. 54-67.

Gillis, Justin. "Antarctic Dispatches." *The New York Times,* 18 May 2017.

Claeys, Gregory. *Dystopia: A Natural History*. Oxford University Press, 2017.

Claeys, Gregory. *Searching for Utopia: The History of an Idea*. Thames & Hudson, 2011.

Motesharrie, S., J. Rivas and E. Kalnay. "Human and nature dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies," *Ecological Economics,* No. 101 (2014), pp. 90–102.

Nietzsche, Friedrich: https://en.wikiquote.org/wiki/Friedrich_Nietzsche

NOAA: U. S., National Oceanic and Atmospheric Administration, *Global National and Regional Sea Level Rise Scenarios for the United States*, January 2017: https://tidesandcurrents.noaa.gov/publications/techrpt83_Global_and_Regional_SLR_Scenarios_for_the_US_final.pdf.

Tainter, Joseph. *The Collapse of Complex Societies*. Cambridge University Press, 1990.

CHAPTER 6:

AI Control Problem: https://en.wikipedia.org/wiki/AI_control_problem

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. "Concrete Problems in AI Safety." (2016, July 25), arXiv16066.06565: https://arxiv.org/abs/1606.06565

Bates, Samantha. "Computers Gone Wild: Impact and Implications of Developments in Artificial Intelligence on Society" (2016, May 9): https://cyber.harvard.edu/node/99484

Blackford, Russell & Damien Broderick (eds.). *Intelligence Unbound: The Future of Uploaded and Machine Minds*. Wiley Blackwell, 2014.

Bostrom, Nick. *Superintelligence*. Oxford University Press, 2014. [See especially Chapter 7, "The Superintelligent Will."]

Brooks, M. "Your quantum brain." *New Scientist*, v. 228, no. 3050 (Dec. 5-11, 2015), pp. 28-31.

Burton, Robert A. "Our A. I. President." *The New York Times*, 22 May 2017.

Chalmers, David. "The Singularity: A Philosophical Analysis" (2010): http://consc.net/papers/singularity.pdf

Dowd, Maureen. "Elon Musk's Billion-Dollar Crusade to Stop the A.I. Apocalypse," *Vanity Fair*, April 2017, P. 116: http://www.vanityfair.com/news/2017/03/elon-musk-billion-dollar-crusade-to-stop-ai-space-x

Existential Risk from Artificial General Intelligence: https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence

"Human, all-too-human": A book title (1878-1880) by Friedrich Nietzsche, *Menshliches, Allzumenshliches*: https://en.wikipedia.org/wiki/Human,_All_Too_Human

Khatchadourian, R. "The Doomsday Invention: Will artificial intelligence destroy us?" *The New Yorker*, 23 November 2015, pp. 64-79.

Kolata, Gina. "Who Needs Hard Drives? Scientists Store Film Clip in DNA." *The New York Times*, 12 July 2017:

https://www.nytimes.com/2017/07/12/science/film-clip-stored-in-dna.html?mcubz=2.

Kurzweil, R. "Book Review: How we'll end up merging with our technology." *The New York Times*, 14 March 2017.

Nicolelis, M. A. L. "Brain-to-Brain Interfaces: When Reality Meets Science Fiction." *Cerebrum*, September 2014: file:///C:/Users/Administrator/Downloads/Brain-to-Brain-Interfaces.pdf

Orseau, L. and S. Armstrong "Safely Interruptible Agents" (2016, June 1): https://intelligence.org/files/Interruptibility.pdf

CHAPTER 7:

Pepper: https://www.ald.softbankrobotics.com/en/cool-robots/pepper

CHAPTER 8:

ALPHAGO: https://en.wikipedia.org/wiki/AlphaGo and https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol

Beauregard, M. & V. Paquette. "Neural correlates of mystical experience in Carmelite nuns." *Neuroscience Letters*, No. 405 (2006), pp. 186–190.

Bodanis, David. *$E=mc^2$: A Biography of the World's Most Famous Equation*. Toronto: Anchor Canada, 2001.

Bostrom, Nick and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence" (2011): http://www.nickbostrom.com/ethics/artificial-intelligence.pdf

British Broadcasting Corporation (2016): http://www.bbc.com/earth/story/20160120-you-decide-what-is-the-most-beautiful-equation-ever-written

Centre for the Study of Existential Risk, University of Cambridge, "Open Letter for Robust AI Research": http://cser.org/open-letter/

Christofori, Irene *et al*. "Neural correlates of mystical experience." *Neuropsychologia*, No. 80 (2016), pp. 212–220.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. "Fairness through Awareness" (2011, November 30), arXiv1104.3913v2: https://arxiv.org/pdf/1104.3913.pdf

Future of Humanity Institute, Strategic Artificial Intelligence Research Centre: https://www.fhi.ox.ac.uk/research/research-areas/strategic-centre-for-artificial-intelligence-policy/

Future of Life Institute: AI Safety Research. https://futureoflife.org/ai-safety-research/

Gopnik, Alison. "4-Year Olds don't act like Trump." *The New York Times*, 20 May 2017. http://www.alisongopnik.com/; https://en.wikipedia.org/wiki/Alison_Gopnik

Greene, Brian. "That Famous Equation and You." *The New York Times,* September 30, 2005*:* http://www.nytimes.com/2005/09/30/opinion/that-famous-equation-and-you.html

Hartnett, Kevin. "How to force our machines to play fair." *Quanta Magazine* (2016, November 23)*:* https://www.quantamagazine.org/20161123-privacy-and-fairness-an-interview-with-cynthia-dwork/

Hubble Space Telescope, picture of Jupiter: https://www.nytimes.com/2017/04/07/science/jupiter-photos-hubble-telescope-juno-nasa.html

Kaku, Michio. "Are we becoming gods?" *New Scientist,* No. 2628 (03 November 2007), pp. 58-9.

Kaku, Michio. *Physics of the Impossible*. New York: Random House, 2008.

McAfee, Andrew and E. Brynjolfsson. "Where Computers Defeat Humans, and Where They Can't." *The New York Times,* March 16, 2016: https://www.nytimes.com/2016/03/16/opinion/where-computers-defeat-humans-and-where-they-cant.html?_r=0 [AlphaGo]

Machine Intelligence Research Institute, "Why AI Safety": https://intelligence.org/why-ai-safety/

Nadella, Satya. "The Partnership of the Future" (2016, June 28): http://www.slate.com/articles/technology/future_tense/2016/06/microsoft_ceo_satya_nadella_humans_and_a_i_can_work_together_to_solve_society.html

Owen, Amy D. *et al*. "Religious Factors and Hippocampal Atrophy in Late Life." *PLoS ONE*, 6(3), 2011, p. e17006: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017006

Nicolelis, Miguel. *Beyond Boundaries: The New Neuroscience of Connecting Brains with Machines---and How It Will Change Our Lives*. New York: Times Books, 2011.

Nicolelis, Miguel & Ronald Cicurel. *The Relativistic Brain: How it Works and Why it cannot be Simulated by a Turing Machine*. CreateSpace International Publishing Platform, 2015. [See especially Chapter 5.]

Partnership on Artificial Intelligence: https://www.partnershiponai.org/

Proctor, D., R. A. Williamson, F. B. M. de Waal, and S. F. Brosnan. "Chimpanzees play the ultimatum game." *Proceedings of the National Academies of*

Sciences, vol. 110, no. 6 (2013, February 3): http://www.pnas.org/content/110/6/2070.abstract

Russell, S., Dewey, D. and M. Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence" (2015, Winter): https://futureoflife.org/data/documents/research_priorities.pdf?x33688

Singer, Tania and Olga M. Klimecki. "Empathy and Compassion." *Current Biology,* Vol. 24, No. 18 (2014), R875-R878: http://www.cell.com/current-biology/references/S0960-9822(14)00770-2

Smolin, Lee. *Time Reborn*. Boston: Houghton Mifflin, 2013.

Wolchover, Natalie. "Concerns of an Artificial Intelligence Pioneer." *Quanta Magazine* (2015, April 21)*:* https://www.quantamagazine.org/20150421-concerns-of-an-artificial-intelligence-pioneer/

Werfel, Franz: https://en.wikipedia.org/wiki/Franz_Werfel

Yourgrau, Palle. *A World without Time: The Forgotten Legacy of Gödel and Einstein*. New York: Basic Books, 2005: http://www.friesian.com/goedel.htm

CHAPTER 10:

Lin, Patrick, Keith Abney & George A. Bekey (eds.). *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, 2012.

http://moralmachine.mit.edu/ ; https://en.wikipedia.org/wiki/Robot_ethics

ten Brinke, Leanne *et al.* "Crocodile Tears: Facial, Verbal and Body Language Behaviours Associated with Genuine and Fabricated Remorse." *Law and Human Behavior*, vol. 36 (2012), pp. 51-9. 9https://people.ok.ubc.ca/stporter/Publications_files/Crocodile%20tears.pdf

Wallach, Wendell & Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, 2008.

Acknowledgements

Professor Richard Smith of Simon Fraser University, former student and treasured colleague, has helped me over many years with trading references, discussing new ideas, pleasant visits with family, encouragement with my sometimes-zany thoughts, managing my website, and countless forms of generous assistance, including now the production and marketing of this book.

I am at the age when old friends and professional colleagues matter more with each passing year, and none more so than Steve Kline of Vancouver and St. Catherines, and Ian Angus, professor at Simon Fraser, another former student and now colleague. Also Steve Hrudey, Marc Saner, Harrie Vredenburg, Jackie Botterill, Ed and Julia Levy, Kyle Asquith, Cameron Pallett, Len Ritter, Greg Paoli, and Dan Krewski. Nick Thorkelson of Boston has involved me in his wonderful project on the life of Herbert Marcuse, which continually brings back fond memories of my long apprenticeship with a great teacher. Many fellows at The Royal Society of Canada have become good friends as well as workers in the common causes associated with national academies around the world.

After our meeting in the Summer of 2013, Alex Colville gave me permission to use three of his paintings as cover artwork for this trilogy. I am most grateful to Alex Colville for this gracious act, and to his daughter Anne Kitz for assisting me with the arrangements for securing the transparencies and permissions from the museums that hold two of these paintings.

In at least some respects, the painting style known as “magic realism” fits well with the minor literary genre of utopian fiction. Both ask their audiences to accept their obvious violations of normal experience as the point of entry into another dimension of existence. More particularly, both seek to achieve their effect on us in the same way, namely, by concealing beneath the work’s surreal surface layer an elaborate, precisely drawn architecture—in the painting, an exact geometry of space, and in the fictional work, a methodical dialogue about ideas. The intended effect is, of course, a conjurer’s trick, the creation of an illusion. By willingly suspending disbelief, the audience can pass through the work’s portal and live for a while in another dimension of space and time.

The short piece entitled “About the Herasaga,” also found in this Back Section, contains an interpretation of the imagery in the three Colville paintings used as cover artwork in this trilogy. No one should regard this brief interpretive exercise as anything else than my own purely idiosyncratic reading of these settings. No claim is made that the painter himself had any such meanings in mind for his creations.

Figure 18 Puncak Jaya (Carstenz Pyramid), Mount Jayawijawa, Papua Province, Indonesia

About The Herasaga

HERA, OR EMPATHY
BOOK TWO: THE PRIESTHOOD OF SCIENCE
BOOK THREE: HERA THE BUDDHA

Thematic Outline:
The Way of Reflection on mind's relation to nature passes through the moments of submission (religion) and dominion (technology) toward its goal—mind's peace with nature.
Since the beginnings of human civilization 6,000 years ago in the Near East—in Egypt and Mesopotamia—Mind (human thinking) has been at war with Nature in two vastly different but complementary forms, namely, religion and technology. In both of these forms, Nature is nothing in itself, simply a background field of matter and energy onto which human meaning and power is projected and imposed.

Represented systematically, this process develops as follows:

1. *Positing:* The religious representation of reality posits Nature passively, not self-originating, as created by Absolute Spirit or God, in which Mind participates derivatively as Soul.
2. *Negation:* In order to fulfill itself as technology, Mind posits Nature as the Other to itself, merely "mindless" matter and energy governed by laws, and in this way finally unlocks the secrets of its own self-origins as Nature (that is, as the product of DNA's evolution).
3. *Negation of negation:* Mind dissolves Spirit and understands itself as natural, as a product of nature, and thus as limited in time, not infinite or absolute.

Book One. *Hera, or Empathy:*

The cover artwork, Colville's "Church and Horse" (1964), shows nature in opposition to the religious representation of reality, in which nature is the passive outcome of an act of creation. Here religion is portrayed as a lifeless empty façade (the building) and broken domain (the gate), and is juxtaposed to the fierce energy and determination of the living, riderless animal that moves menacingly toward the standpoint of the viewer, unstoppably away and out.

Book Two. *The Priesthood of Science:*

The cover artwork, Colville's "Horse and Train" (1954), shows human technology in its head-on confrontation with a vastly more powerful, living nature. The horse, again riderless and uncontrolled, moving away from the viewer, paradoxically towering in size over the train and opposing itself fearlessly to it, sets itself squarely upon the tracks, eschewing the surrounding fields.

Book Three. *Hera the Buddha:*

The cover artwork, Colville's "Moon and Cow" (1963), shows a much-domesticated animal—which is thus itself both a human creation yet still also living nature—at rest in the night, after a long day of her labors in the service of human needs (but with no human masters now present), facing away from the viewer, at peace with nature.

[1] The case of Fritz Haber (1868–1934) illustrates many of the themes under discussion here. Born a Jew in Breslau (now Wroclaw), Haber converted to Lutheranism and became a strong German nationalist. He won the Nobel Prize in Chemistry in 1918 for developing the synthesis of ammonia from nitrogen and hydrogen; the Haber-Bosch Process created industrial quantities of synthetic fertilizer, and it is estimated that half of the human population is alive today due to its impact on agriculture and food production. (It also became a key component in high-explosive shells and bombs.) During World War I his nationalistic fervor prompted him to seek an officer's commission, then to weaponize chlorine gas and personally supervise its release against enemy soldiers on both the Western and Eastern fronts. In the early 1920s scientists working in Haber's lab at the Kaiser Wilhelm Institute developed a cyanide-based pesticide commercialized as Zyklon A; its successor, Zyklon B, was used to kill millions in the Nazi extermination camps. At the University of Berlin Haber and Einstein became close personal friends despite their sharply differing political views, but Haber dutifully voted to terminate Einstein's membership in the Prussian Academy after Einstein fled Germany in 1932. He himself lost his university professorship in 1933, despite Max Planck's attempt to intervene on his behalf directly with Hitler by recalling Haber's service to the German state (Hitler told Planck, "A Jew is a Jew"), and he died in exile in Switzerland the following year.

[2] A wonderful treasure from this period is the *Born–Einstein Letters*. Einstein's close friend Max Born (1882– 1970) was, like Haber and Immerwahr, a Jew born in Breslau (Wroclaw) and became one of the great geniuses of atomic physics during the 1920s, known among other achievements as one of the founders of quantum mechanics. Like so many others he left Germany after 1933 and became a professor at Edinburgh before returning to Germany in his retirement. He and Einstein never saw each other again after both emigrated, but thankfully their extraordinary wartime correspondence survives.