The Herasaga

Book One: Hera, or Empathy

Book Two: The Priesthood of Science

Book Three: Hera the Buddha

Note to the Reader:

This PDF file contains Chapter 6, "The Threat of Superintelligence," in the book entitled *Hera The Buddha*; the entire volume is available as an E-book on Amazon, as follows:

https://www.amazon.ca/Hera-Buddha-Utopian-Fiction-Herasaga-ebook/dp/B074KP7Q1R/ref=sr_1_1?s=books&ie=UTF8&qid=1501945489&sr=1-1&keywords=hera+the+buddha
Product Details:
Author: William Leiss (www.leiss.ca)
Length: Pages xix, 225
Language: English
ISBN 978-0-9738283-2-0

Please use the following citation when quoting from this Material:

> William Leiss, "The Threat of Superintelligence," Chapter 6 in *Hera The Buddha: A Work of Utopian Fiction* (Book 3 of *The Herasaga*), Kindle Digital Publishing, 2017.

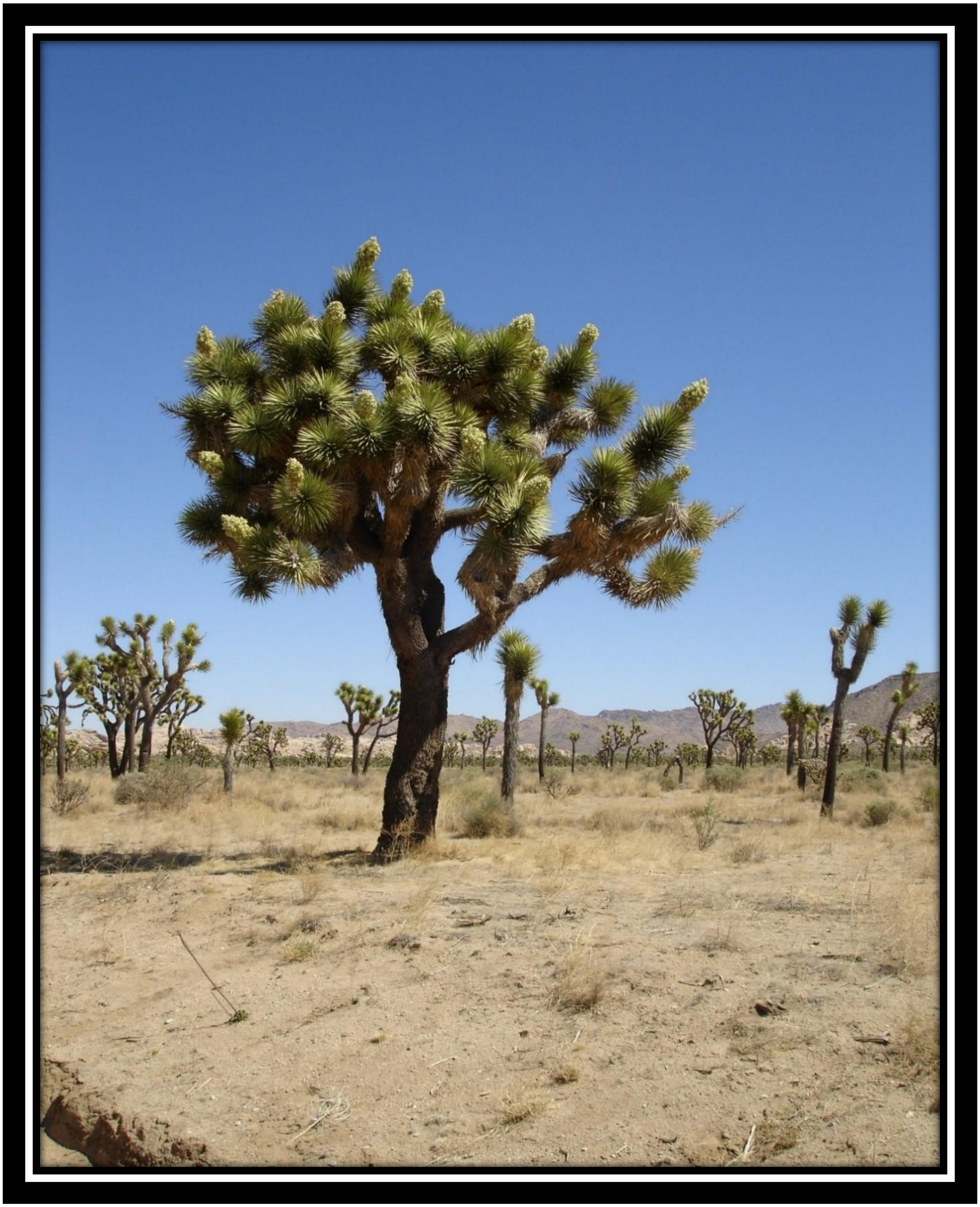

Figure 1 Yucca brevifolia in bloom, Joshua Tree National Park, California (Photo: W. Leiss)

HERA THE BUDDHA

A WORK OF UTOPIAN FICTION

WILLIAM LEISS

$$e^{i\pi}+1=0$$

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

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?

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

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