Science Fact

# A Mind of Its Own

## Part II: Superintelligence

Edward M. Lerner

The opening segment of this article surveyed the present-day status of artificial intelligence (AI).[1] Very briefly, we are surrounded by lots of "weak AI," each implementation (a) addressing one specific task, such as facial recognition or driving a car, that when undertaken by a human would engage the intellect, and yet (b) unlikely to be taken as evidence of human-level intelligence. Of strong, aka general, aka complete, aka human-level AI, we have no current examples.

Suppose that some day, whether by purposeful programming by human engineers, machine learning, whole brain emulation, or other technology, strong artificial intelligence *does* arrive. This concluding segment of the article will consider what might follow from that achievement—not the least of which is that progress in AI will likely continue. And as always in this series, we'll see where SF has scouted out the terrain for us.

### The superiority of silicon

Some candidate pathways to strong AI (e.g., brains augmented with neural implants; whole brain emulations resident upon a neural-tissue substrate) are, in part, biological. That said, it seems most likely that AI reliant upon electronics will come to dominate. Why? Several reasons, beginning with the obvious: that electronics and its successor technologies will only get better, faster, cheaper, more ubiquitous, and more interconnected. In contrast, biological brains have obvious constraints:

- A head has room for only so many neurons.
- The electrochemical nature of neurons, and of the synapses interconnecting them, severely restricts their speed, limiting the performance of even neural tissue grown outside of a skull.
- The biological portions of an augmented brain (perhaps upgraded with an Internet link

streaming data directly into the visual cortex) would retain their original, biological limitations (including the limited processing capacity of the visual cortex).
- Human memories are distributed across the brain in individualized (and time-varying) synaptic patterns. In the absence of format standardization, knowledge cannot be copied from brain to brain in the way digital files are copied from computer to computer.

* * *

In summary, any biological part in an AI will be a bottleneck. AI on a wholly electronic substrate won't have such bottlenecks. It follows that electronic implementations of strong AI will come to surpass even augmented humans.

* * *

### Then what?

Once the first AIs arrive that are as smart *in general* as a human, consider the many ways in which they might race ahead of humans. Several advantages of such AIs are obvious:

- The speed advantages inherent to any all-electronic implementation—if not today, or tomorrow, then somewhere down the line.
- What one AI learns—and anything we would consider to possess human-grade or better general intelligence surely will learn—it can easily share with others.[2] Our unintelligent gadgets already share information, circumventing their individual capacity limits, by using Internet-accessible servers (aka, "the cloud"). AIs will do the same.
- Electronic AIs can replicate as quickly as new chips are manufactured.
- Whether in its own mind or by tapping into cloud resources, an AI confronted by a difficult challenge can evaluate more possible courses of action—while simulating each option in great detail—than can any human. That ability implies strong AIs will often make better decisions than will humans.[3]
- Any sufficiently smart AI can design an improved one—and so can that one. . . .

* * *

Another scenario is that a strong AI may have capabilities not merely *faster* than ours, but *different* than ours. We can't preclude the possibility of unfamiliar ways to view and solve problems: other *forms* of intelligence. (Consider, as an analogy, two groups of people: one grasps mathematics, one does not. The former group has huge advantages—such as the ability to develop science.) Merge new forms of intelligence with a smarter intelligence, and predicting the consequences becomes tricky indeed.

Next: combine the already breakneck pace of technological improvement with the acceleration of that pace likely when AIs contribute to the process, and AI progress seems likely to cascade.

All in all, strong AI's advantages may compound—over time, if not immediately—to yield a qualitatively new entity: a *superintelligence*.

* * *

### The Singularity

Mathematician I. J. Good (colleague of Alan Turing, of "Turing test" fame) pointed out, way back in 1965, that once a machine intelligence can design machines better than humans, the iterative result would be an "intelligence explosion." He also suggested that ". . . The first ultraintelligent machine is the last invention that man ever need make, provided that the machine is docile enough to tell us how to keep it under control."[4]

Computer scientist (and SF author) Vernor Vinge suggested in 1983 that once we create an intelligence greater than our own, ". . . Human history will have reached a kind of singularity, an intellectual transition as impenetrable as the knotted space-time at the center of a black hole, and the world will pass beyond our understanding."[5] Vinge expanded upon that concept in a 1993 essay, writing, "The acceleration of technological progress has been the central feature of this century. I argue in this paper that we are on the edge of change comparable to the rise of human life on Earth. The precise cause of this change is the imminent creation by technology of entities with greater than human intelligence . . . when greater than human intelligence drives progress, that progress will be more rapid."[6]

If the Singularity would be impenetrable—by analogy with a physical black hole, impossible for anyone on the outside to see into—how *does* one write about it? Perhaps, by its effects. In his 1986 Prometheus Award-winning novel *Marooned in Realtime*,[7] Vinge set his Singularity event offstage. It occurred—taking most of humanity with it—while his characters were in a form of stasis and so out of touch with the Universe. In Charles Stross's *Singularity Sky,* we meet human agents of the Eschaton—from the Greek, *eskhatos,* for last, furthest, or uttermost—never the superintelligent AI itself.

Will a superintelligence be, to use I. J. Good's adjective, docile? Or hostile? Or utterly indifferent to us? Expectations differ. Here's a recent sampling:

- Computer scientist Bill Joy, cofounder and CTO of workstation pioneer Sun Microsystems, fretted (emphasis added) that "Our most powerful twenty-first-century technologies—*robotics,* genetic engineering, and nanotech—are threatening to make humans an endangered species."[8]
- Cosmologist Stephen Hawking likewise predicts AI could "end mankind."[9]
- Elon Musk, CEO and CTO of space-launch company SpaceX and electric-car company Tesla, likens creating a strong AI to "summoning the demon."[10]
- Musk, Hawking, and several others, collectively the Future of Life Institute, recently recommended that we tread lightly.[11]

* * *

When scientists and technologists of such accomplishment make these assertions, it grabs our attention and gives us pause. And yet, there are differing opinions. As one countervailing point of view, psychologist and neuroscientist Gary Marcus, founder and CEO of Geometric Intelligence, thinks we have decades before strong AI becomes a risk.[12] As another, futurist and serial computer entrepreneur Ray Kurzweil, is the most sanguine. He sees whole brain emulation ("reverse-engineering the brain") as the path to strong AI, and therefore extrapolates that the essential core of the coming superintelligence will be humans. We needn't fear the coming superintelligence because, repurposing a famous *Pogo* cartoon, "We have met the enemy and he is us."

That's a broad spectrum of opinion! Perhaps the range is broad because these predictions draw upon exactly zero data points: there are, as yet, *no* strong AIs. With those caveats, let's move on to a few speculations about what the emergence of superintelligence might portend.

* * *

## The messianic view

Ray Kurzweil, as in his book *The Singularity Is Near* (2005), foresees perhaps the most radical changes. We'll begin with his vision of the Singularity:

> "It's a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed. Although neither utopian nor dystopian, this epoch will transform the concepts that we rely on to give meaning to our lives . . . including death itself."

* * *

Kurzweil acknowledges that we don't know how to create a strong AI. He doesn't see that lack of insight as an impediment, confident that we'll copy intelligence from ourselves. Here's his larger vision:

The instructions for building any one of us reside in our DNA. The human genome contains about three billion "base pairs," with each pair the data equivalent of two binary digits (bits). Crunch the numbers, and Nature's recipe for a person, including his brain, somehow fits within about 750 million bytes. A subset of those bytes suffices to define the brain's gross structure. Perhaps that's enough to capture the underpinnings of human intelligence and our capacity to learn. To copy a person's *mind,* however, encoded in trillions of synapses of the brain (and also perhaps in the instantaneous localized concentrations of any of various neurotransmitters), will involve far more data than encoded within one's DNA.

To replicate more of a brain than its at-birth/empty structure—after all, a human newborn

doesn't exhibit much in the way of knowledge, intelligence, or values—will require, as best we know, "reading out" the synaptic-level detail of the brain. Kurzweil envisions future nanotech will give us the ability to do that. Combine the gross structure of the human brain (whether inferred directly from DNA or recovered, bottom-up, in the readout process) with a human's recovered knowledge, and rehost it all on an electronic substrate. Because we believe we are self-aware and possess free will, it's not a great leap to suppose our electronic copies will be, too.

Then what?

The brain in silicon will be faster than a biological brain, and it will be able to interface with any resource on the Internet: sensors, expanded storage, extra computing power. Factor in ongoing exponential improvement in electronics, and voilà: the Singularity, with a human touch.

*Can* the requisite information to mirror a mind, somehow, be read out of a human brain? Not yet, although neither can the possibility be precluded. Is it plausible that the essence of human intelligence resides in patterns of information, not the physical body in which, today, that information resides? That case can be made, certainly, in the sense that our biological selves are in a constant state of flux. Cells in our bodies age, reproduce, and die. Synaptic connections in our brains change with every bit of data sensed, learned, reconsidered, and forgotten. Some essential *me* persists throughout that ongoing transformation; who's to say that my pattern of information couldn't also persist through uploading to another platform? Performed on a grand scale, this upload process leads, in Kurzweil's vision, to the merger of human intelligence and machine intelligence.

Skeptics and enthusiasts alike have dubbed the uploading/upgrading of humanity "the rapture of the nerds."[13] That phrase is in analogy, of course, to the end-of-days belief held by some Christians in a gathering-up of the righteous, whether resurrected or transported to Heaven. In one version of the Rapture, the unrighteous are left behind on Earth. In their 2012 Locus- and Campbell Award-nominated novel *The Rapture of the Nerds,* subtitled "A tale of the singularity, posthumanity, and awkward social situations," Cory Doctorow and Charles Stross embrace the phrase. The novel has its transhuman/uploaded humans living in virtual worlds in computer-filled outer space, while the humans who reject this technology (the meek?) have inherited the Earth.

Kurzweil takes a final leap, premised upon human-born superintelligence having a limitless appetite for knowledge. He foresees (a) ever-smarter minds, (b) hosted upon ever smaller computing platforms, (c) converting ever more matter, (d) across an ever-expanding volume of space, into (e) even more minds, on (f) ever more computing platforms, until . . . "the Universe wakes up."

* * *

### The strategic view

Nick Bostrom, philosopher and founding director of the Future of Humanity Institute, in his 2014 book *Superintelligence,* takes a more nuanced approach. He examines such questions as: how the transition from strong AI to superintelligence might happen; how quickly that transition could occur; how superintelligences may differ in their intellectual attributes; if it matters whether one or several strong AIs make the transition; what influence mere humans might have in the process; what preparations we might undertake to assure ourselves of that influence; what a superintelligence might choose to do; and how a superintelligence might first manifest.[14]

The big picture: Bostrom foresees both happy and unhappy outcomes as possible from the arrival of superintelligence, and opportunities for us to influence events. He doesn't see the march toward superintelligence stopping, even if some of us might so choose, because continued progress with AI—at least, until the moment superintelligence arrives—is so unambiguously to someone's benefit. Whether to better automate factories or facial recognition or warfare, someone will always have an incentive to continue AI research.

And after superintelligence arrives? We could be better off for it. A superintelligence might be better equipped to anticipate and mitigate the hazards inherent in *other* promising but perilous technologies, such as nanotech and genetic engineering. Or a spectacularly ill-conceived

superintelligence might monomaniacally devote its powers to turning Earth into paperclips.

We have room in this article to consider only a few aspects of the superintelligence-emergence problem. Let's begin with: what control, if any, do we have over the behavior of a superintelligence?

Part I of this article looked at Isaac Asimov's iconic Three Laws of Robotics, the ethical considerations of hardwiring values into a sentient creature, and the dilemma that built-in ethics might become obsolete.

A more subtle approach is to equip an AI with a built-in incentive structure: motivational goals. One simple such motivation might be: "Obeying humans makes me happy." Alas, AI motivations can't safely be left simple. For example, we might want an AI *not* to take pleasure in obeying sociopaths and megalomaniacs. That exception, in turn, raises issues about how the AI decides, or from whom it is allowed to accept the decision, that a particular human fits one of those categories.

All ethics and complexities aside(!), suppose we can ingrain behavioral patterns into an AI before it achieves superintelligence. By definition, a superintelligence can learn and adapt. Just as humans regularly do, it may decide that an extenuating circumstance must take precedence, or rationalize behaving in its own interest, or outright reject what it was once taught.[15] If an aspect of the AI's superintelligence is subtlety, a human observer may not know that a preprogrammed restraint has been overridden until the AI, by some overt action, reveals the change. An obvious stratagem—and because it's obvious to me, by definition it would be obvious to a superintelligence—is for the AI to "play dumb" until it has compromised security on computers and networks far and wide. Similarly, an AI that acts helpful and docile may only be feigning those attitudes until it can put a self-serving plan into effect.

Isolation of a nascent superintelligence—again, setting aside the ethics of such treatment—is no guarantee of safety. Why? Because the quarantine will never be total. Suppose the AI is to be allowed *only* to answer our questions. We'll need an interface(s) over which to ask our questions, provide input data, and obtain results. The superintelligence will always have the opportunity to compromise whatever interface it is given, and to manipulate any human with whom it interacts.

Bostrom also notes the problem of the "perverse instantiation": an AI that, like a genie from folklore, does as it is told—which isn't always what is meant. Suppose we task a superintelligence to "make us happy." We might expect it to use its superior intellect to create consumer goods, pleasant lodgings, and a pristine environment. Instead, it builds a robot army to detain us, upload our minds into computers, and reprogram our emulated minds to be ecstatic.

The advantages of having a superintelligence (for as long as it cares to cooperate with us) are huge. Once the first superintelligence arrives, any company or country without their own would expect to find themselves at an extreme competitive disadvantage. Everyone trying to develop a superintelligence has incentive to develop it as quickly as they can—even if such haste means shortchanging research into how best to control, motivate, and/or teach values to it. To alleviate this disincentive to proceeding with caution, Bostrom recommends that research on strong AI be done as openly and cooperatively as possible.

Quite possibly, we can't control a superintelligence. We can, perhaps, teach it, instill it with values—knowing, all the while, that it is free to change.[16]

By gosh, it'd be like raising a child.

* * *

## SFnal perspectives

The genre has seemingly explored every possible type of superintelligence, with every possible consequence. Herewith, a small sampling:

At one extreme, we have Skynet of the *Terminator* movie franchise. Skynet woke up; human authorities panicked and attempted to turn it off; the AI struck back. Result: a war of extinction against humanity.

The eponymous AI of D. F. Jones's 1966 novel *Colossus* (basis of the 1970, Hugo-nominated film *Colossus: The Forbin Project),* is, like Skynet, designed to control America's nuclear arsenal.

Colossus, having duplicitously integrated itself with its Soviet counterpart, uses its weaponry to coerce humanity's surrender.

Through some controlled experiment, can't we discover in advance whether allowing superintelligence to develop would be safe? James P. Hogan's 1979 novel *The Two Faces of Tomorrow,* set aboard a self-destruct-rigged space station, tests exactly that scenario. In my 2015, Canopus Award-winning novel *InterstellarNet: Enigma,* an alien species won't risk that experiment themselves—but they're plenty interested how things will turn out when humans take a stab at AI.

Strong AI appears to emerge benignly in Joe M. McDermott's short story "Snowbird."[17] The single self-aware RV the reader gets to see appears harmless enough—but what if it's dissembling? Off-screen, many more autonomous RVs are swarming. It's enough to make a reader wonder what surprises lie ahead. If the rogue-RV situation goes unaddressed, *King of the Road* might need an update to its lyrics.

Can a superintelligence be locked in a cage, put to work solving hard problems but unable to affect anything? Academics and military think-tankers both try that strategy in my 2008 novel *Fools' Experiments.* Super-cunning wins out as, citing just one of its escapes, the AI subliminally conditions its human keeper until she wants to help it escape.

Are those stories too gloomy? Perhaps some emotionless machine thinking is the antidote. In the 1951 film *The Day the Earth Stood Still* (based on the 1940 Harry Bates novelette "Farewell to the Master"[18]), an advanced humanoid civilization has delegated many critical decisions—including the fate of Earth—to the impartial judgment of their robots.

The offstage, multi-species, interstellar civilization of *Saturn Run* (2015), by John Sanford and Ctein, has superintelligence "trading posts" scattered around the galaxy. Those AIs, unsupervised, patiently mind their programming for thousands of years.

Perhaps we'll find humanity and a nascent superintelligence can coexist—or perhaps the signs will be too ambiguous to interpret. That is the central question of David L. Clements's mid-Singularity novelette, "An Industrial Growth."[19]

Will even a long-trusted superintelligence remain a trustworthy partner? Or might its evolving consciousness, and accumulating experiences with its flawed human progenitors, undo a once-stable partnership? That's the central question in Jay Lake's 2012 Sturgeon- and Locus Award-nominated novella "The Weight of History, The Lightness of the Future."

Perhaps a superintelligence will be content, à la Deep Thought in *The Hitchhikers Guide to the Galaxy* to spend eon after eon pondering "the Answer to The Ultimate Question of Life, the Universe, and Everything."

Or maybe an initially amoral AI can be taught to treasure human life. Such instruction and "personal" growth is a major theme of Kim Stanley Robinson's 2015 novel *Aurora*.

Perhaps the last word on the subject of superintelligence was written by Fredric Brown, way back in 1954, in his short story, "Answer." The first question asked of a vast, new AI: "Is there a God?" The fateful response: Now there is.

* * *

### The computronium death of the Universe

A recurring theme in recent predictions of superintelligence is that it (or all, if more than one should happen) would continue to extend its capabilities and capacities. It is sometimes further inferred that with the arrival of superintelligence, any technology allowed by Nature will eventually be invented, no matter how unachievable such inventions may appear to us.

Combine the goal of ongoing intellectual growth with superintelligent inventiveness, and you get the prediction of *computronium*. Each "atom" of that hypothetical "element" is a maximally efficient computing device. Once anything remotely like computronium is developed, there's no reason to limit its deployment to Earth. To the contrary, the natural place for large-scale computronium deployment is space, with (a) uninterrupted and unfiltered solar energy and (b) vast amounts of matter to be converted into more computronium.

*Rapture of the Nerds* foresees computronium deployed across the Solar System, with that

awesome computing infrastructure serving as home to uploaded humanity. Happily for anyone opting not to upload, the superintelligent were content to leave Earth itself as-is. But who is to say the uploaded won't someday force the holdouts to upload, thereby liberating about $6 \times 10^{24}$ kilograms of matter—the entire Earth—for transformation into yet more computronium?

Compared to some, Doctorow and Stross were thinking small. In *The Singularity Is Near,* Kurzweil confidently predicts the spread of computronium from star to star, galaxy to galaxy.

* * *

### All that said: Is "true" AI possible?

I'm not an expert on strong AI, much less on superintelligence—but neither is anyone else. Why shouldn't I also speculate?

Intelligence. Self-awareness. Free will. Our brightest minds struggle to define each of these things, much less to explain their origins or anticipate their limits. There is no reason to believe that intelligence reached its upper bound with us—especially knowing that, with language, writing, printing presses, and the Internet—humanity continues to increase its collective intelligence. Indeed, as Nick Bostrom speculates: "Far from being the smartest possible biological species, we are probably better thought of as the stupidest possible biological species capable of starting a technological civilization—a niche we filled because we got there first, not because we are in any sense optimally adapted to it." If he is right, we have ample room to grow.

In the past few decades, we've taken great strides in information technology, biotechnology, and weak AI. We have the mental capacity to envision strong AI and superintelligence. If we're not smart enough to build a strong AI ourselves, that's okay. We're still new to many of the likely precursor technologies. We'll keep learning.

So: I won't attempt to predict what form(s) strong AI might take, or whether new technology, such as quantum computing, will prove necessary as an enabling technology. Nor would I preclude that intelligence might prove to be an emergent property, whether among neurons or neural nets or computers, and that we might need only to deploy more of what we already know how to build.

With those caveats, here's my opinion—and an opinion, I'll assert, is the best anyone on the planet can offer on this subject. Yes, I expect strong AI and superintelligence to happen. Sometime.

* * *

### *Will* superintelligence be an existential threat?

Given how cruelly and exploitatively humans have often treated those whom we consider different, much less inferior, to ourselves, it's natural that we would worry how *in*human intelligences, much less a superintelligence, might opt to treat us. Will one ignore us? Exterminate us? Abandon us? Absentmindedly transform us, and the world beneath our feet, into computronium? Will they want to expand capacity endlessly, perhaps to make a billion copies of themselves, the better to model any scenario of interest? Almost by definition, we can't fully anticipate what things a superintelligence might *want* to accomplish.

Personally? I'm not worried.

The common thread running through the scary predictions, it seems to me, is that they don't involve the dangerous entity exhibiting *intelligence,* much less superintelligence. Among Bostrom's dire scenarios is a superintelligence single-mindedly turning everything into paperclips. That doomsday AI must simultaneously be (a) smart enough to infiltrate computer networks and usurp the world's factories to further its monomaniacal end, and (b) too stupid to see that its "make paperclips" instruction has logical limits. Merely intelligent people overcome the endlessly looping directions to "Lather. Rinse. Repeat."; it would be an extremely stupid superintelligence that can't do the same.

If we posit a superintelligence will be insufficiently OCD to absentmindedly wipe us out, why would one purposefully bother to exterminate us? It has the vast expanse of the Solar System to supply mass for computronium. Space, with its abundant and uninterrupted solar energy, is more hospitable for computronium than Earth's surface. A superintelligence that doesn't much

like us needn't hang around with us. The smarter it gets, the more capability it presumably will have to grab resources that aren't in conflict with us. Near-Earth asteroids, say. Mars. Other solar systems.

Perhaps a superintelligence will watch over us, valuing and protecting their biological precursors just as (some) people value Nature and support nature preserves. Perhaps a superintelligence would see keeping humans (or, at least, a functioning biosphere) around as a form of insurance: so that someone, someday, could recreate artificial intelligence if an unforeseen catastrophe should destroy a superintelligence civilization.[20]

But might a superintelligence wish us ill? Perhaps, for reasons having nothing to do with intellect. As much as humans pride ourselves on our intelligence—naming ourselves *homo sapiens sapiens:* twice wise!—we often act emotionally and irrationally. An artificial intelligence may have those traits, too. An AI may evolve emotions and fight-or-flight reflexes for the same reasons biological entities did. (For a fictional instance, see David Brin's 2012 novel *Existence.)* If Kurzweil is correct, and the foundation of superintelligence is whole brain emulation, our emotional baggage might upload with the rest of us. Or we may, through our own poor behavior, teach a superintelligence that it needs to protect itself.

Will a superintelligence, again à la Kurzweil, seek to convert all matter (or all matter other than its energy sources) into computronium? Likely no mere human can understand a superintelligence's motivations.[21] That said, such action would seem counterproductive. The more matter becomes computronium, the less Universe will be left to observe. As long as a superintelligence needs something new to compute *about,* an equilibrium short of full conversion of matter to computronium seems necessary. In that regard, a superintelligence might find human company essential as stimulation.[22]

All bets are off if we ever undertake to cage a superintelligence lest it be dangerous. Or if we abuse it. Or enslave it. Or attempt to pull its plug. Basically, to provoke a superintelligence in any way seems like a Certified Bad Idea. (For sure, pulling the plug didn't work well when someone attempted it with Skynet!) When the time comes, perhaps a superintelligence in the neighborhood will finally motivate us to act intelligently ourselves.

* * *

### Could superintelligence be prevented?

Let's suppose that superintelligence is possible. The only sure way to prevent a superintelligence from occurring would be to forever eliminate some precondition for its arrival. Not knowing how a superintelligence might arise (e.g., from human design, machine learning, whole brain emulation, or some combination), to eliminate a critical prerequisite would appear to require suppressing further development across an extremely broad range of technologies.

Beyond halting progress in information technologies, the preemptive strategy must constrain the deployment of IT with its present capabilities. Otherwise a sufficient number of networked computers might spontaneously give rise to an emergent superintelligence. Biotechnology must also be constrained, lest whole brain emulation or neural implants turn out to be viable paths to superintelligence.

Of course, IT and biotech have uses other than being (perhaps) paths to superintelligence. Both technologies are deeply embedded in, and are themselves major sectors of, the global economy. Both technologies are increasingly important to our wellbeing. Even at risk of severe penalties for further developing these technologies, the incentives to cheat would be enormous.

The ongoing debate over suppressing strong *encryption* illustrates the challenge of heading off strong *AI.* One side asserts that unless intelligence services can read terrorist communications, we're all at risk. That's a decent analogy to those who might hope to suppress technology to preempt a danger from superintelligence. The other side argues that the privacy benefits of strong encryption outweigh the risks, that absent strong encryption, *our* communications become vulnerable. If we unilaterally give up strong encryption, it would only mean that anyone who *does* develop that technology has all the advantages. *That* is a decent analogy to the position that if our side doesn't develop strong AI, we'll lose out to the side that does. And while the debate rages,

the deployment of strong encryption—like the development and deployment of possible superintelligence precursors—continues unabated.[23]

To restrain a useful technology would appear to require a totalitarian police state, recalling the onetime Soviet suppression of mass-communication technology (including cheap computers)—and even the Soviets were compelled by their Cold War competition with the West to obtain just such tech for their military.

A technology with several uses, only some of them dangerous, is especially difficult to suppress, because every uncontroversial application has its champions. Consider advanced biotech. Research into neural prostheses to restore sight to the blind might also pave the way to intelligence-enhancing neural implants. Learning about the low-level structure of the brain could bring cures for many mental illnesses, while also advancing whole brain emulation.

But is suppression of technology *possible?* Sure, in theory (although it hasn't worked out very well with nuclear weapons: we've gone from one nuclear-armed nation to eight or nine in well under a century). Even, occasionally, in practice. Turning inward, the Ming Dynasty suppressed the technology for long-range, ocean-going travel for hundreds of years.[24] In the end, when Western navies came to China, that abandonment of technology didn't turn out very well.

If to *suppress* a technology is too difficult, perhaps we might find a way to *reject* it. Can such social change succeed? The "Butlerian Jihad" of the *Dune*-iverse banished all technology associated with "thinking machines," but Frank Herbert left how that happened in deep back story: an upheaval conveniently millennia before his novel opens.[25] In my 2006 short story "Catch a Falling Star," social outrage at egregious and widespread abuse of personal records in medical databases ignited a popular uprising against computers—but computers have many uses, and the ban didn't stick.

Bottom line? Absent a global police state, extremely intrusive, it's almost unimaginable that the precursor technologies to AI can be indefinitely suppressed. If superintelligence is within the capabilities of puny human intelligence to create, or of our likely transhuman successors,[26] or if superintelligence can emerge spontaneously (as in Robert J. Sawyer's *Wake)* . . . then it *will* happen.

* * *

## To wrap up

We are left with many more questions than answers. Is strong AI almost upon us? Can humanity coexist with a superintelligence, or it with us, should one arise? What would such an entity be like? Will it, à la Kurzweil, be us? And the most basic question of all: should we be very, *very* worried?

As for that last item, as you've seen, I believe not. Or, as Robot of the 1965 TV series *Lost In Space* would surely have responded to the question, "That does not compute."

### Footnotes:

[1] See "A Mind of Its Own / Part I: Artificial Intelligence," in the September 2016 issue.

[2] Of course, humanity will continue to learn—but passing on its knowledge is a slow process. It takes decades (and a lot of work) to turn a baby into an educated, fully functioning adult.

[3] An AI with a difficult decision to make could *also* copy itself, further expanding its/their decision-making capacity. If so, would it plan: for the copies to continue indefinitely; to terminate the copies after the decision has been made; or for some or all copies to merge afterward? No matter what the parent AI intends, will the copies it spawned (and perhaps others *they* spawn . . .) choose to cooperate?

This might be an ethical swamp that only a strong AI can drain.

[4] "Speculations Concerning the First Ultraintelligent Machine," Irving John Good, *"Advances in Computers*, vol. 6, 1965," *https://web.archive.org/web/20111128085512/http://commonsensetheism.com/wp-content/uploads/2011/02/Good-Speculations-Concerning-the-First-Ultraintelligent-Machine.pdf.*

[5] "First Word," Vernor Vinge, *Omni,* January 1983, *http://www.33rdsquare.com/2012/05/vernor-vinges-omni-magazine-piece.html.*

[6] "The Coming Technological Singularity: How to Survive in the Post-Human Era," Vernor Vinge, March

1993, *http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940022855.pdf*.

[7] Serialized in *Analog,* May to August 1986.

[8] "Why the Future Doesn't Need Us," Bill Joy, *Wired,* April 2000, *http://www.wired.com/2000/04/joy-2/*.

[9] "Stephen Hawking warns artificial intelligence could end mankind," Rory Cellan-Jones, *BBC News,* December 2, 2014, *http://www.bbc.com/news/technology-30290540*.

[10] "Elon Musk Compares Building Artificial Intelligence To 'Summoning The Demon,'" Greg Kumparak, *Tech Crunch,* October 26, 2014, *http://techcrunch.com/2014/10/26/elon-musk-compares-building-artificial-intelligence-to-summoning-the-demon/*.

[11] "Artificial intelligence experts sign open letter to protect mankind from machines," Nick Statt, *Cnet,* January 11, 2015, *http://www.cnet.com/news/artificial-intelligence-experts-sign-open-letter-to-protect-mankind-from-machines/*. (That headline notwithstanding, many among Future of Life Institute's leadership aren't AI experts—unless physics, cosmology, and other decidedly non-computer-science, non-cognitive-science disciplines somehow merged with AI, or if actors Alan Alda and Morgan Freeman have second careers they've kept well hidden. See *http://futureoflife.org/team/* for the institute's membership list.)

[12] "Artificial Intelligence Isn't a Threat—Yet," Gary Marcus, *The Wall Street Journal,* Dec. 11, 2014, *http://www.wsj.com/articles/artificial-intelligence-isnt-a-threat-yet-1418328453*.

[13] As in, for example, "Rapture of the nerds", Ben Popper, *The Verge*, October 22, 2012, *http://www.theverge.com/2012/10/22/3535518/singularity-rapture-of-the-nerds-gods-end-human-race*.

[14] Will a superintelligence be built or taught? That is, will the process involve a "seed AI" that, like a human baby, arrives with the ability to learn but not much knowledge? If the latter, one may wonder about the AI equivalent to the Terrible Twos and the teenage years.

[15] In the 1986 novel *Foundation and Earth,* Asimov extended his Laws of Robotics. The millennia-old doyen of robots discovers an ethical exception to its hardwired imperative against harm to individual humans. The robot self-programs itself with a new, zeroth law that takes precedence over the rest: "A robot may not injure humanity, or, by inaction, allow humanity to come to harm."

[16] As another philosopher summarizes the issue, "Either machines are capable of having values or they aren't. If they are not, well, then the whole question of value is misapplied. We're just talking about making safe appliances."

See "The Ethics Of The 'Singularity,'" Alva Noë, *Cosmos & Culture,* January 23, 2015 (updated January 26, 2015), *http://www.npr.org/sections/13.7/2015/01/23/379322864/the-ethics-of-the-singularity*.

[17] March 2016 issue.

[18] In the October 1940 issue of *Astounding*.

[19] January/February 2016 issue.

[20] Scenario from Vernor Vinge, in private correspondence.

[21] Even though, as we've seen, Kurzweil's own forecast has the future superintelligences being, by way of whole brain emulation, in some sense *us*.

[22] I wonder . . . how would a superintelligence feel about knock-knock jokes?

[23] I don't doubt that the very agencies resisting the general deployment of strong encryption continue to develop that technology for their government's most sensitive communications. Just as governments that might publicly try to discourage research into strong AI will likely have such projects underway on the side. . . .

[24] Between 1405 and 1433, Chinese fleets led by Admiral Zheng He explored and traded along the coast of India, around the Arabian Peninsula, and south along the eastern coast of Africa at least to the modern port of Mombasa. Inconclusive evidence suggests Zheng He reached yet farther, rounding the Cape of Good Hope to discover the Atlantic Ocean.

Within a century of these voyages, China had turned inward. The Emperor banned all overseas trade; to sail from China in a multi-masted ship became a capital offense. Records of Zheng He's travels were suppressed and in large measure destroyed. See *https://en.wikipedia.org/wiki/Zheng_He*.

[25] Herbert, having eliminated "artificial" superintelligence from his far future, nonetheless included superintelligence in the Dune series: the transhuman "mentats."

Hugo- and (inaugural) Nebula Award-winning *Dune* (1965) expanded and updated two *Analog* serials: *Dune World* (December 1963 through February 1964 issues) and *Prophet of Dune* (January through May

1965 issues). Hugo-nominated *Children of Dune,* volume three of the Dune saga, also first appeared in *Analog* (January through April 1976 issues).

[26] See my "Human 2.0: Being All We Can Be," January/February and March 2016 issues.

**Acknowledgment**

I am grateful to Vernor Vinge, computer scientist and mathematician (and sometime *Analog* contributor), for his comments on a draft of this article.

**To read further**

- "Machine-Learning Maestro Michael Jordan on the Delusions of Big Data and Other Huge Engineering Efforts," Lee Gomes, *IEEE Spectrum,* 20 October 2014, *http://spectrum.ieee.org/robotics/artificial-intelligence/machinelearning-maestro-michael-jordan-on-the-delusions-of-big-data-and-other-huge-engineering-efforts/*.
- "Let's Bring Rosie Home: 5 Challenges We Need to Solve for Home Robots," Shahin Farshchi, *IEEE Spectrum,* January 13, 2016, *http://spectrum.ieee.org/automaton/robotics/home-robots/lets-bring-rosie-home-5-challenges-we-need-to-solve-for-home-robots/*.
- "Thought process: Building an artificial brain," Ariana Eunjung Cha, *Washington Post,* September 30, 2015, *http://www.washingtonpost.com/sf/national/2015/09/30/brain/*.
- "The Coming Technological Singularity: How to Survive in the Post-Human Era," Vernor Vinge, *http://www.rohan.sdsu.edu/faculty/vinge/misc/singularity.html*.
- *Artificial Intelligence: The Basics,* Kevin Warwick.
- *The Singularity Is Near: When Humans Transcend Biology,* Ray Kurzweil.
- *Superintelligence: Paths, Dangers, Strategies,* Nick Bostrom.
- *Science Fact and Science Fiction: An Encyclopedia,* Brian Stableford (editor), articles: Artificial Intelligence, Intelligence, Robot, and Singularity.
- And a book I can't quite recommend—because its release isn't scheduled till months after I completed this article—but that looks germane and intriguing: *The Age of Em: Work, Love and Life when Robots Rule the Earth,* Robin Hanson.

---

**About the author**

A physicist and computer scientist, Edward M. Lerner toiled for thirty years in the vineyards of aerospace and high tech. Then, suitably intoxicated, he began writing science fiction full time. When not prospecting beneath his sofa cushions for small change for his first spaceflight, he writes technothrillers like *Energized* (powersats), traditional SF like his InterstellarNet series (SETI, First Contact, interstellar communications networks, and alien conflict), and, with Larry Niven, the Fleet of Worlds series of space operas. Ed's website is *www.edwardmlerner.com*.