Are We Enlightened Guardians, or Are We Apes Designing Humans?

Douglas Mulhall*

Author, "Our Molecular Future" & "The Calcium Bomb"

Most students of artificial intelligence are familiar with this forecast made by Vernor Vinge in 1993:[1] "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended."

That was thirteen years ago. Many proponents of super-intelligence say we are on track for that deadline, due to the rate of computing and software advances. Skeptics argue this is nonsense and that we're still decades away from it.

But fewer and fewer argue that it won't happen by the end of this century. This is because history has shown the acceleration of technology to be exponential, as explained in well-known works by inventors such as Ray Kurzweil and Hans Moravec, some of which are elucidated in this volume of essays.

A classic example of technology acceleration is the mapping of the human genome, which achieved most of its progress in the late stages of a multi-year project that critics wrongly predicted would take decades. The rate of mapping at the end of the project was exponential compared to the beginning, due to rapid automation that has since transformed the biotechnology industry.

The same may be true of molecular manufacturing (MM) as self-taught machines learn via algorithms to do things faster, better, and cheaper. I won't describe the technology of MM here because that is well covered in other essays by more competent experts.

MM is important to super-intelligence because it will revolutionize the processes required to understand our own intelligence, such as neural mapping via neural probes that non-destructively map the brain. It also will accelerate three-dimensional computing, where the space between computing units is reduced and efficiency multiplied in the same way that our own brains have done it. Once this happens, the ability to mimic the human brain will accelerate, and self-aware intelligence may follow quickly.

* E-mail: mulhall@calcify.com.

[1] The Coming Technological Singularity: How to Survive in the Post-Human Era http://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.

Nanotechnology Perceptions **2** (2006) 161–166
Received 24 April 2006

This type of acceleration suggests that Vinge's countdown to the beginning of the end of the human era must be taken seriously.

The pathways by which super-human intelligence could evolve have been well explained by others and include: computer-based artificial intelligence, bioelectronic AI that develops super-intelligence on its own, or human intelligence that is accelerated or merged with AI. Such intelligence might be an enhancement of *Homo sapiens*, i.e. part of us, or completely separate from us, or both.

Many experts argue that each of these forms of super-intelligence will enhance humans, not replace them, and although they might seem alien to unenhanced humans, they will still be an extension of us because we are the ones who designed them.

The thought behind this is that we will go on as a species.

Critics, however, point to a fly in that ointment. If the acceleration of computing and software continues apace, then super-intelligence, once it emerges, could outpace *Homo sapiens*, with or without piggybacking on human intelligence.

This would see the emergence of a new species, perhaps similar in some ways, but in other ways fundamentally different from *Homo sapiens* in terms of intelligence, genetics, and immunology.

If that happens, the gap between *Homo sapiens* and super-intelligence could quickly become as wide as the gap between apes and *Homo sapiens*.

Optimists say this won't happen, because everybody will get an upgrade simultaneously when super-intelligence breaks out.

Pessimists say that just a few humans or computers will acquire such intelligence first, and then use it to subjugate the rest of us *Homo sapiens*.

For clues as to who might be right, let's look at outstanding historical examples of how we've used technology and our own immunology in relation to less technologically adept societies, and in relation to other species.

When technologically superior Europeans arrived in North and South America, the indigenous populations didn't have much time to contemplate such implications because in a just few years, most who came in contact with Europeans were dead from disease. Many who died never laid eyes on a European, as death spread so quickly ahead of the conquerors through unknowing victims.

Europeans at first had no idea that their own immunity to disease would give them such an advantage, but when they realized it, they did everything to use it as a weapon. They did the same with technologies that they consciously invented and knew were superior.

The rapid death of these ancient civilizations, numbering in the tens of millions of persons across two continents, is not etched into the consciousness of contemporary society because those cultures left few written records and had scant time to document their own demise. Most of what they put to pictures or symbols was destroyed by religious zealots or wealth-seeking exploiters.

And so, these civilizations passed quietly into history for the most part.

By inference, enhanced intelligence easily could take choices about our future out of our hands, and may also be immune to hazards such as mutating viruses that pose dire threats to human society.

Annihilation of *Homo sapiens* could occur in one of many ways:

- The "oops" factor: accidental annihilation at the hands of a very smart klutz, e.g. by something that is unwittingly immune to things that kill us, or that is smart in one way, but inept in others. Predecessors to super-intelligence may only be smarter than us in *some* ways, and therein lies a danger. An autistic intelligence could do us in by accident. Just look at current technology, where computers are more capable than humans in some ways but hopeless in others.
- Annihilation in the crossfire of a war-like competition between competing forms of super-intelligence, some of which might include upgraded *Homo sapiens*. One of the early, deadlier competitions could be for resources as various forms of super-intelligence gobble up space that we occupy, or remake our ecology into an environment more suitable to their needs.
- Deliberate annihilation or assimilation because we are deemed inferior.

If Vernor Vinge is right, we have 18 years before we will face such realities. Centuries ago, the fate of Indian civilizations in North and South America was decided in a similar time span. So, the time to address such risks is now.

This is especially true because paradigms shift more quickly now; therefore, when the event occurs we'll have less time, perhaps five years or even just one, to consider our options.

What might we use as protection against these multi-factorial threats?

Sun Microsystems' cofounder Bill Joy's April 2000 treatise, "Why the future doesn't need us",[2] summarized one field of thought, arguing the case for relinquishment— eschewing certain technologies due to their inherent risks.

Since that time, most technology proponents have been arguing why relinquishment is impractical. They contend that the march of technology is relentless and we might as well go along for the ride, but with safeguards built in to make sure things don't get too crazy.

Nonetheless, just how we build safeguards into something smarter than us, including an upgraded version of ourselves, has as yet gone unanswered. To see where the solutions might lie, let's again look at the historical perspective.

If we evaluate the arguments between technology optimists and relinquishment pessimists in relation to the history of the natural world, it becomes apparent that we are stuck between a rock and a hard place.

The 'rock' in this case could be an asteroid or comet. If we were to relinquish our powerful new technologies, chances are good that an asteroid would eventually collide with Earth, as has occurred before, thus throwing human civilization back to the dark ages or worse.

For those who scoff at this as an astronomical long shot, be reminded that Comet Shoemaker-Levy 9 punched Earth-sized holes in Jupiter less than a decade after the space tools necessary to witness such events were launched, and just when most experts were forecasting such occurrences to be once-in-a-million-year events that we would likely never see.

Or perhaps we would be thrown back by other catastrophic events that have occurred historically, such as naturally-induced climate changes triggered by super-volcanos, collapse of the magnetosphere, or an all-encompassing supernova.

Due to those natural risks, I argue in my book, *Our Molecular Future*, that we may have no choice but to proceed with technologies that could just as easily destroy us as protect us.

[2] http://www.wired.com/wired/archive/8.04/joy.html.

Unfortunately, as explained in the same book, an equally bad 'hard place' sits opposite the onrushing "rock" that threatens us. The hard place is our social ineptness.

In the 21st century, despite tremendous progress, we still do amazingly stupid things. We prepare poorly for known threats including hurricanes and tsunamis. We go to war over outdated energy sources such as oil, and some of us increasingly overfeed ourselves while hundreds of millions of people ironically starve. We often value conspicuous consumption over saving impoverished human lives, as low-income victims of AIDS or malaria know too well.

Techno-optimists use compelling evidence to argue that we are vanquishing these shortcomings and that new technologies will overcome them completely. But one historical trend bodes against this: *emergence of advanced technologies has been overwhelmingly bad for many of the less intelligent species on Earth.*

To cite a familiar refrain: We are massacring millions of wild animals and destroying their habitat. We keep billions more domestic farm animals under inhumane, painful, plague-breeding conditions in increasingly vast numbers.

The depth and breadth of this suffering is so vast that we often ignore it, perhaps because it is too terrible to contemplate. When it gets too bothersome, we dismiss it as animal rights extremism. Some of us rationalize it by arguing that nature has always extinguished species, so we are only fulfilling that natural role.

But at its core lies a searing truth: our behavior as guardians of less intelligent species, which we know feel pain and suffering, has been and continues to be atrocious.

If this is our attitude toward less intelligent species, why would the attitude of superior intelligence toward us be different? It would be foolish to assume that a more advanced intelligence than our own, whether advanced in all or in only some ways, will behave benevolently toward us once it sees how we treat other species.

We therefore must consider that a real near-term risk to our civilization is that we invent something that looks at our ways of treating less intelligent species and decides we're not worth keeping, or if we are worth keeping, we should be placed in zoos in small numbers where we can't do more harm. Resulting questions:

- How do we instill into super-intelligence 'ethical' behavior that we ourselves poorly exhibit?
- How do we make sure that super-intelligence rejects certain unsavory practices as we banned slavery?
- Can we reach into the future to prevent a super-intelligence from changing its mind about those ethics?

These questions have been debated, but no broad-based consensus has emerged. Instead, as the discussions run increasingly in circles, they suggest that we as a species might be comparable to 'apes designing humans'.

The ape-like ancestors of *Homo sapiens* had no idea they were contributing DNA to a more intelligent species. Nor could they hope to comprehend it. Likewise, can we *Homo sapiens* expect to comprehend what we are contributing to a super-intelligent species that follows us?

As long as we continue to exercise callous neglect as guardians of species less intelligent than ourselves, it could be argued that we are much like our pre-human ancestors: incapable of consciously influencing what comes after us.

The guardianship issue leads to another question: How well are we balancing technology advantages against risks?

In the mere 60 years since our most powerful weapons—nuclear bombs—were invented, we've kept them mostly under wraps and congratulated ourselves for that, but we have also seen them proliferate from at first just one country to at least ten, with some of those balanced on the edge of chaos.

Likewise, in the nanoscale technology world that precedes molecular manufacturing, we've begun assessing risks posed to human health by engineered nanoparticles, but those particles are already being put into our environment and into us.

In other words, we are still closing the proverbial barn doors after the animals have escaped. This limited level of foresight is light years away from being able to assess how to control the onrushing risks of molecular manufacturing or of enhanced intelligence.

Many accomplished experts have pointed out that the same empowerment of individuals by technologies such as the Internet and biotech could make unprecedented weapons available to small disaffected groups.

Technology optimists argue that this has occurred often in history: new technologies bring new pros and cons, and after we make some awful mistakes with them, things get sorted out.

However, in this case the acceleration rate by its nature puts these technologies in a class of their own, because the evidence suggests they are running ahead of our capacities to contain or balance them. Moreover, the number of violently disaffected groups in our society who could use them is substantial.

To control this, do we need a "pre-crime" capacity as envisaged in the film *Minority Report*, where Big Brother methods are applied to anticipate crime and strike it down preemptively?

The pros and cons of preemptive strikes have been well elucidated recently. The idea of giving up our freedom in order to preserve our freedom from attack by disaffected groups is being heavily debated right now, without much agreement.

However, one thing seems to have been under-emphasized in these security debates:

Until we do the blatantly positive things such as eliminate widespread diseases, feed the starving, house the homeless, disenfranchise dictators, stop torture, stop inhumane treatment of less intelligent species, and other do-good things that are treated today like platitudes, we will not get rid of violently disaffected groups.

By doing things that are blatantly humane, (despite the efforts of despots and their extremist anti-terrorist counterparts to belittle them as wimpy) we might accomplish two things at once: greatly reduce the numbers of violently disaffected groups, and present ourselves to super-intelligence as being enlightened guardians.

Otherwise, if we continue along the present path, we may someday seem to super-intelligence what our ape-like ancestors seem to us: primitive.

In deciding what to do about *Homo sapiens*, a superior form of intelligence might first evaluate our record as guardians, such as how we treat species less intelligent than ourselves, and how we treat members of our same species that are less technologically adept or just less fortunate.

Why might super-intelligences look at this first? Because just as we are guardians of those less intelligent or fortunate than us, so super-intelligences will be the guardians of us and of other less intelligent species. Super-intelligences will have to decide what to do with us, and with them.

If Vinge is accurate in his forecast, we don't have much time to set these things straight before someone or something superior to us makes a harsh evaluation.

Being nice to dumb animals or poor people is by no means the only way of assuring survival of our species in the face of something more intelligent than us. Using technology to massively upgrade human intelligence is also a prerequisite. But that, on its own, may not be sufficient.

Compassion by those who possess overwhelming advantages over others is one of the special characteristics that *Homo sapiens* (along with a few other mammals) brings to this cold universe. It is what separates us from an asteroid or supernova that doesn't care whether it wipes us out.

Further, compassionate behavior is something most of us could agree on, and while it is often misinterpreted by some as a weakness, it is also what makes us human, and what most of us would want to contribute to future species.

If that is so, then let's take the risk of being compassionate and put it into practice by launching overarching works that demonstrate the best of what we are.

For example, use molecular manufacturing and its predecessor nanotechnologies to eliminate the disease of aging, instead of treating the symptoms. That is what I personally have decided to focus on, but there are many other good examples out there, including synthesized meat that eliminates inhumane treatment of billions of animals, and cheap photovoltaic electricity that could slash our dependence on oil—and end wars over it.

Such works are not hard to identify. We just have to give them priority. Perhaps then we will seem less like our unwitting ancestors and more like enlightened guardians.

About the author:

Douglas Mulhall is the author of *Our Molecular Future: How Nanotechnology, Robotics, Genetics*, and *Artificial Intelligence Will Transform Our World*, and co-author of *The Calcium Bomb: The Nanobacteria Link to Heart Disease and Cancer*. He managed a scientific environmental institute for many years and co-founded one of the early South American institutes devoted to recycling technology.