Why the future doesn't need us.

Our most powerful 21st-century technologies - robotics, genetic engineering, and nanotech - are threatening to make humans an endangered species.

By Bill Joy

From the moment I became involved in the creation of new technologies, their ethical dimensions have concerned me, but it was only in the autumn of 1998 that I became anxiously aware of how great are the dangers facing us in the 21st century. I can date the onset of my unease to the day I met Ray Kurzweil, the deservedly famous inventor of the first reading machine for the blind and many other amazing things.

Ray and I were both speakers at George Gilder's Telecosm conference, and I encountered him by chance in the bar of the hotel after both our sessions were over. I was sitting with John Searle, a Berkeley philosopher who studies consciousness. While we were talking, Ray approached and a conversation began, the subject of which haunts me to this day.

I had missed Ray's talk and the subsequent panel that Ray and John had been on, and they now picked right up where they'd left off, with Ray saying that the rate of improvement of technology was going to accelerate and that we were going to become robots or fuse with robots or something like that, and John countering that this couldn't happen, because the robots couldn't be conscious.

While I had heard such talk before, I had always felt sentient robots were in the realm of science fiction. But now, from someone I respected, I was hearing a strong argument that they were a near-term possibility. I was taken aback, especially given Ray's proven ability to imagine and create the future. I already knew that new technologies like genetic engineering and nanotechnology were giving us the power to remake the world, but a realistic and imminent scenario for intelligent robots surprised me.

It's easy to get jaded about such breakthroughs. We hear in the news almost every day of some kind of technological or scientific advance. Yet this was no ordinary prediction. In the hotel bar, Ray gave me a partial preprint of his then-forthcoming book The Age of Spiritual Machines, which outlined a utopia he foresaw - one in which humans gained near immortality by becoming one with robotic technology. On reading it, my sense of unease only intensified; I felt sure he had to be understating the dangers, understating the probability of a bad outcome along this path.

I found myself most troubled by a passage detailing adystopian scenario:

THE NEW Luddite CHALLENGE

First let us postulate that the computer scientists succeed in developing intelligent machines that
can do all things better than human beings can do them. In that case presumably all work will be
done by vast, highly organized systems of machines and no human effort will be necessary. Either of
two cases might occur. The machines might be permitted to make all of their own decisions without
human oversight, or else human control over the machines might be retained.

If the machines are permitted to make all their own decisions, we can't make any conjectures as to
the results, because it is impossible to guess how such machines might behave. We only point out
that the fate of the human race would be at the mercy of the machines. It might be argued that the
human race would never be foolish enough to hand over all the power to the machines. But we are
suggesting neither that the human race would voluntarily turn power over to the machines nor that
the machines would willfully seize power. What we do suggest is that the human race might easily
permit itself to drift into a position of such dependence on the machines that it would have no
practical choice but to accept all of the machines' decisions. As society and the problems that face it
become more and more complex and machines become more and more intelligent, people will let
machines make more of their decisions for them, simply because machine-made decisions will bring
better results than man-made ones. Eventually a stage may be reached at which the decisions
necessary to keep the system running will be so complex that human beings will be incapable of
making them intelligently. At that stage the machines will be in effective control. People won't be
able to just turn the machines off, because they will be so dependent on them that turning them off
would amount to suicide.

On the other hand it is possible that human control over the machines may be retained. In that case
the average man may have control over certain private machines of his own, such as his car or his
personal computer, but control over large systems of machines will be in the hands of a tiny elite
just as it is today, but with two differences. Due to improved techniques the elite will have greater
control over the masses; and because human work will no longer be necessary the masses will be
superfluous, a useless burden on the system. If the elite is ruthless they may simply decide to
exterminate the mass of humanity. If they are humane they may use propaganda or other
psychological or biological techniques to reduce the birth rate until the mass of humanity becomes
extinct, leaving the world to the elite. Or, if the elite consists of soft-hearted liberals, they may
decide to play the role of good shepherds to the rest of the human race. They will see to it that
everyone's physical needs are satisfied, that all children are raised under psychologically hygienic
conditions, that everyone has a wholesome hobby to keep him busy, and that anyone who may
become dissatisfied undergoes "treatment" to cure his "problem." Of course, life will be so
purposeless that people will have to be biologically or psychologically engineered either to remove
their need for the power process or make them "sublimate" their drive for power into some harmless
hobby. These engineered human beings may be happy in such a society, but they will most certainly
not be free. They will have been reduced to the status of domestic animals.¹

In the book, you don't discover until you turn the page that the author of this passage is Theodore
Kaczynski - the Unabomber. I am no apologist for Kaczynski. His bombs killed three people during a
17-year terror campaign and wounded many others. One of his bombs gravely injured my friend
David Gelernter, one of the most brilliant and visionary computer scientists of our time. Like many of
my colleagues, I felt that I could easily have been the Unabomber's next target.

Kaczynski's actions were murderous and, in my view, criminally insane. He is clearly a Luddite, but
simply saying this does not dismiss his argument; as difficult as it is for me to acknowledge, I saw
some merit in the reasoning in this single passage. I felt compelled to confront it.

Kaczynski's dystopian vision describes unintended consequences, a well-known problem with the
design and use of technology, and one that is clearly related to Murphy's law - "Anything that can go
wrong, will." (Actually, this is Finagle's law, which in itself shows that Finagle was right.) Our
overuse of antibiotics has led to what may be the biggest such problem so far: the emergence of
antibiotic-resistant and much more dangerous bacteria. Similar things happened when attempts to
eliminate malarial mosquitoes using DDT caused them to acquire DDT resistance; malarial parasites
likewise acquired multi-drug-resistant genes.²

The cause of many such surprises seems clear: The systems involved are complex, involving
interaction among and feedback between many parts. Any changes to such a system will cascade in
ways that are difficult to predict; this is especially true when human actions are involved.
I started showing friends the Kaczynski quote from *The Age of Spiritual Machines*: I would hand them Kurzweil's book, let them read the quote, and then watch their reaction as they discovered who had written it. At around the same time, I found Hans Moravec's book *Robot: Mere Machine to Transcendent Mind*. Moravec is one of the leaders in robotics research, and was a founder of the world's largest robotics research program, at Carnegie Mellon University. *Robot* gave me more material to try out on my friends - material surprisingly supportive of Kaczynski's argument. For example:

**The Short Run (Early 2000s)**

Biological species almost never survive encounters with superior competitors. Ten million years ago, South and North America were separated by a sunken Panama isthmus. South America, like Australia today, was populated by marsupial mammals, including pouched equivalents of rats, deers, and tigers. When the isthmus connecting North and South America rose, it took only a few thousand years for the northern placental species, with slightly more effective metabolisms and reproductive and nervous systems, to displace and eliminate almost all the southern marsupials.

In a completely free marketplace, superior robots would surely affect humans as North American placentals affected South American marsupials (and as humans have affected countless species). Robotic industries would compete vigorously among themselves for matter, energy, and space, incidentally driving their price beyond human reach. Unable to afford the necessities of life, biological humans would be squeezed out of existence.

There is probably some breathing room, because we do not live in a completely free marketplace. Government coerces nonmarket behavior, especially by collecting taxes. Judiciously applied, governmental coercion could support human populations in high style on the fruits of robot labor, perhaps for a long while.

A textbook dystopia - and Moravec is just getting wound up. He goes on to discuss how our main job in the 21st century will be "ensuring continued cooperation from the robot industries" by passing laws decreeing that they be "nice," and to describe how seriously dangerous a human can be "once transformed into an unbounded superintelligent robot." Moravec's view is that the robots will eventually succeed us - that humans clearly face extinction.

I decided it was time to talk to my friend Danny Hillis. Danny became famous as the cofounder of Thinking Machines Corporation, which built a very powerful parallel supercomputer. Despite my current job title of Chief Scientist at Sun Microsystems, I am more a computer architect than a scientist, and I respect Danny's knowledge of the information and physical sciences more than that of any other single person I know. Danny is also a highly regarded futurist who thinks long-term - four years ago he started the Long Now Foundation, which is building a clock designed to last 10,000 years, in an attempt to draw attention to the pitifully short attention span of our society. (See "Test of Time," *Wired* 8.03, page 78.)

So I flew to Los Angeles for the express purpose of having dinner with Danny and his wife, Pati. I went through my now-familiar routine, trotting out the ideas and passages that I found so disturbing. Danny's answer - directed specifically at Kurzweil's scenario of humans merging with robots - came swiftly, and quite surprised me. He said, simply, that the changes would come gradually, and that we would get used to them.

But I guess I wasn't totally surprised. I had seen a quote from Danny in Kurzweil's book in which he said, "I'm as fond of my body as anyone, but if I can be 200 with a body of silicon, I'll take it." It seemed that he was at peace with this process and its attendant risks, while I was not.

While talking and thinking about Kurzweil, Kaczynski, and Moravec, I suddenly remembered a novel I had read almost 20 years ago - *The White Plague*, by Frank Herbert - in which a molecular biologist is driven insane by the senseless murder of his family. To seek revenge he constructs and disseminates a new and highly contagious plague that kills widely but selectively. (We're lucky Kaczynski was a mathematician, not a molecular biologist.) I was also reminded of the Borg of *Star Trek*, a hive of partly biological, partly robotic creatures with a strong destructive streak. Borg-like disasters are a staple of science fiction, so why hadn't I been more concerned about such robotic
dystopias earlier? Why weren't other people more concerned about these nightmarish scenarios?

Part of the answer certainly lies in our attitude toward the new - in our bias toward instant familiarity and unquestioning acceptance. Accustomed to living with almost routine scientific breakthroughs, we have yet to come to terms with the fact that the most compelling 21st-century technologies - robotics, genetic engineering, and nanotechnology - pose a different threat than the technologies that have come before. Specifically, robots, engineered organisms, and nanobots share a dangerous amplifying factor: They can self-replicate. A bomb is blown up only once - but one bot can become many, and quickly get out of control.

Much of my work over the past 25 years has been on computer networking, where the sending and receiving of messages creates the opportunity for out-of-control replication. But while replication in a computer or a computer network can be a nuisance, at worst it disables a machine or takes down a network or network service. Uncontrolled self-replication in these newer technologies runs a much greater risk: a risk of substantial damage in the physical world.

Each of these technologies also offers untold promise: The vision of near immortality that Kurzweil sees in his robot dreams drives us forward; genetic engineering may soon provide treatments, if not outright cures, for most diseases; and nanotechnology and nanomedicine can address yet more ills. Together they could significantly extend our average life span and improve the quality of our lives. Yet, with each of these technologies, a sequence of small, individually sensible advances leads to an accumulation of great power and, concomitantly, great danger.

What was different in the 20th century? Certainly, the technologies underlying the weapons of mass destruction (WMD) - nuclear, biological, and chemical (NBC) - were powerful, and the weapons an enormous threat. But building nuclear weapons required, at least for a time, access to both rare - indeed, effectively unavailable - raw materials and highly protected information; biological and chemical weapons programs also tended to require large-scale activities.

The 21st-century technologies - genetics, nanotechnology, and robotics (GNR) - are so powerful that they can spawn whole new classes of accidents and abuses. Most dangerously, for the first time, these accidents and abuses are widely within the reach of individuals or small groups. They will not require large facilities or rare raw materials. Knowledge alone will enable the use of them.

Thus we have the possibility not just of weapons of mass destruction but of knowledge-enabled mass destruction (KMD), this destructiveness hugely amplified by the power of self-replication.

I think it is no exaggeration to say we are on the cusp of the further perfection of extreme evil, an evil whose possibility spreads well beyond that which weapons of mass destruction bequeathed to the nation-states, on to a surprising and terrible empowerment of extreme individuals.

Nothing about the way I got involved with computers suggested to me that I was going to be facing these kinds of issues.

My life has been driven by a deep need to ask questions and find answers. When I was 3, I was already reading, so my father took me to the elementary school, where I sat on the principal's lap and read him a story. I started school early, later skipped a grade, and escaped into books - I was incredibly motivated to learn. I asked lots of questions, often driving adults to distraction.

As a teenager I was very interested in science and technology. I wanted to be a ham radio operator but didn't have the money to buy the equipment. Ham radio was the Internet of its time: very addictive, and quite solitary. Money issues aside, my mother put her foot down - I was not to be a ham; I was antisocial enough already.

I may not have had many close friends, but I was awash in ideas. By high school, I had discovered the great science fiction writers. I remember especially Heinlein's 'Have Spacesuit Will Travel' and Asimov's 'I, Robot', with its Three Laws of Robotics. I was enchanted by the descriptions of space travel, and wanted to have a telescope to look at the stars; since I had no money to buy or make
become enormously lucrative within this decade, causing a large incremental investment in all nanotechnologies.

Unfortunately, as with nuclear technology, it is far easier to create destructive uses for nanotechnology than constructive ones. Nanotechnology has clear military and terrorist uses, and you need not be suicidal to release a massively destructive nanotechnological device—such devices can be built to be selectively destructive, affecting, for example, only a certain geographical area or a group of people who are genetically distinct.

An immediate consequence of the Faustian bargain in obtaining the great power of nanotechnology is that we run a grave risk—the risk that we might destroy the biosphere on which all life depends.

As Drexler explained:

"Plants" with "leaves" no more efficient than today's solar cells could out-compete real plants, crowding the biosphere with an inedible foliage. Tough omnivorous "bacteria" could out-compete real bacteria: They could spread like blowing pollen, replicate swiftly, and reduce the biosphere to dust in a matter of days. Dangerous replicators could easily be too tough, small, and rapidly spreading to stop—at least if we make no preparation. We have
trouble enough controlling viruses and fruit flies.

Among the cognoscenti of nanotechnology, this threat has become known as the "gray goo problem." Though masses of uncontrolled replicators need not be gray or gooey, the term "gray goo" emphasizes that replicators able to obliterate life might be less inspiring than a single species of crabgrass. They might be superior in an evolutionary sense, but this need not make them valuable.

The gray goo threat makes one thing perfectly clear: We cannot afford certain kinds of accidents with replicating assemblers.

Gray goo would surely be a depressing ending to our human adventure on Earth, far worse than mere fire or ice, and one that could stem from a simple laboratory accident.6 Oops.

It is most of all the power of destructive self-replication in genetics, nanotechnology, and robotics (GNR) that should give us pause. Self-replication is the modus operandi of genetic engineering, which uses the machinery of the cell to replicate its designs, and the prime danger underlying gray goo in nanotechnology. Stories of run-amok robots like the Borg, replicating or mutating to escape from the ethical constraints imposed on them by their creators, are well established in our science fiction books and movies. It is even possible
that self-replication may be more fundamental than we thought, and hence harder—or even impossible—to control. A recent article by Stuart Kauffman in Nature titled "Self-Replication: Even Peptides Do It" discusses the discovery that a 32-amino-acid peptide can "autocatalyse its own synthesis." We don't know how widespread this ability is, but Kauffman notes that it may hint at "a route to self-reproducing molecular systems on a basis far wider than Watson-Crick base-pairing."

In truth, we have had in hand for years clear warnings of the dangers inherent in widespread knowledge of GNR technologies—of the possibility of knowledge alone enabling mass destruction. But these warnings haven’t been widely publicized; the public discussions have been clearly inadequate. There is no profit in publicizing the dangers.

The nuclear, biological, and chemical (NBC) technologies used in 20th-century weapons of mass destruction were and are largely military, developed in government laboratories. In sharp contrast, the 21st-century GNR technologies have clear commercial uses and are being developed almost exclusively by corporate enterprises. In this age of triumphant commercialism, technology—with science as its handmaiden—is delivering a series of almost magical inventions that are the most phenomenally lucrative ever seen. We are
aggressively pursuing the promises of these new technologies within the now-unchallenged system of global capitalism and its manifold financial incentives and competitive pressures.

This is the first moment in the history of our planet when any species, by its own voluntary actions, has become a danger to itself—as well as to vast numbers of others.

It might be a familiar progression, transpiring on many worlds—a planet, newly formed, placidly revolves around its star; life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges which, at least up to a point, confers enormous survival value; and then technology is invented. It dawns on them that there are such things as laws of Nature, that these laws can be revealed by experiment, and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others, not so lucky or so prudent, perish.

That is Carl Sagan, writing in 1994, in *Pale Blue Dot*, a book describing his vision of the human future in space. I am only now realizing how deep his insight was, and how sorely I miss, and will miss, his voice. For all its eloquence,