Nanotechnology Dangers and Defenses

Ray Kurzweil[*]

Kurzweil Technologies, Inc.

The first half of the 21st century will be characterized by three overlapping revolutions—in Genetics, Nanotechnology, and Robotics (GNR). The deeply intertwined promise and peril of these technologies has led some serious thinkers to propose that we go very cautiously, possibly even to abandon them altogether.

A few years ago, computer maven Bill Joy wrote, "We are being propelled into a new century with no plan, no control, no brakes… The only realistic alternative I see is relinquishment: to limit the development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge."[1]

Joy's deep concern about the future grew out of a conversation we had in 1998 about these emerging technologies, and an early draft of *The Age of Spiritual Machines* that I gave him. Although I have a reputation as a technology optimist, it turns out that at public discussions of 'promise and peril', I often spend much of my time defending Joy's position on the feasibility of the dangers that concern him. Indeed, Joy and I agree on both promise and peril.

Technology has always been a mixed blessing, bringing us benefits such as longer and healthier lifespans, freedom from physical and mental drudgery, and many novel creative possibilities on the one hand, while introducing new dangers. Technology empowers both our creative and destructive natures.

Broad relinquishment is contrary to economic progress and is ethically unjustified given the opportunity to alleviate disease, overcome poverty, and clean up the environment. Joy and I also agree that relinquishment of major fields such as genetics ('G'), nanotechnology ('N'), or strong AI / robotics ('R') is not the answer. There is, however, a growing movement advocating exactly that. Bill McKibben, the environmentalist who first brought global warming to our attention, argues in his book *Enough* that we have had 'enough' technology and should not pursue more. However, regulations on safety—essentially fine-grained relinquishment—will remain an appropriate strategy. In that spirit, Joy and I recently wrote a joint op ed piece ("Recipe for Destruction") published in the New York Times on October 17, 2005 criticizing the publication of the 1918 flu genome on the web.

[*] E-mail: raymond@kurzweiltech.com.

[1] http://www.wired.com/wired/archive/8.04/joy_pr.html.

Received 23 January 2006

Dangers to defend against

As technology accelerates toward the full realization of GNR, we will see interweaving potentials: a feast of creativity resulting from human intelligence expanded manifold, combined with many grave new dangers. A quintessential concern that has received considerable attention is unrestrained nanobot replication. Early proposals for molecular manufacturing required trillions of intelligently designed devices to be useful. To scale up to such levels it would have been necessary to enable them to self-replicate, essentially the same approach used in the biological world (that's how one fertilized egg cell becomes the trillions of cells in a human).

Although the self-replication can be hidden and blocked in a variety of ways (for example, Ralph Merkle's proposal[2] for a 'broadcast architecture' in which each replicating entity needs to get the replicating codes from a secure server), the overall system will have self-replication at some level. And in the same way that biological self-replication gone awry (that is, cancer) results in biological destruction, a defect in a mechanism curtailing nanobot self-replication—the so-called gray goo scenario—would endanger all physical entities, biological or otherwise.

Modern proposals, such as the use of large integrated manufacturing systems rather than trillions of quasi-independent nanobots, appear to prevent inadvertent release of destructive self-replication, but in general these safeguards can be worked around by a determined adversary. We see a similar situation today in biological technologies. The ethical guidelines for gene modification technologies adopted at the Asilomar Conference have worked well for over a quarter of a century, but these guidelines would not restrict a would-be bioterrorist because they don't have to follow the guidelines (they don't have to put their 'inventions' through the FDA either).

These guidelines and strategies are likely to be effective for preventing accidental release of dangerous self-replicating nanotechnology entities. But dealing with the intentional design and release of such entities is a more complex and challenging problem. A sufficiently determined and destructive opponent could possibly defeat each of these layers of protections. Take, for example, the broadcast architecture. When properly designed, each entity is unable to replicate without first obtaining replication codes, which are not repeated from one replication generation to the next. However, a modification to such a design could bypass the destruction of the replication codes and thereby pass them on to the next generation. To counteract that possibility it has been recommended that the memory for the replication codes be limited to only a subset of the full code. However, this guideline could be defeated by expanding the size of the memory.

Another protection that has been suggested is to encrypt the codes and build in protections in the decryption systems, such as time-expiration limitations. However, we can see how easy it has been to defeat protections against unauthorized replications of intellectual property such as music files. Once replication codes and protective layers are stripped away, the information can be replicated without these restrictions.

This doesn't mean that protection is impossible. Rather, each level of protection will work only to a certain level of sophistication. The meta lesson here is that we will need to place twenty-first-century society's highest priority on the continuing advance of defensive

[2] "Self replicating systems and low cost manufacturing" (1994) http://www.zyvex.com/nanotech/selfRepNATO.html.

technologies, keeping them one or more steps ahead of the destructive technologies (or at least no more than a quick step behind).

Living creatures—including humans—would be the primary victims of an exponentially spreading nanobot attack. The principal designs for nanobot construction use carbon as a primary building block. Because of carbon's unique ability to form four-way bonds, it is an ideal building block for molecular assemblies. Because biology has made the same use of carbon, pathological nanobots would find the Earth's biomass an ideal source of this primary ingredient.

How long would it take an out-of-control replicating nanobot to destroy the Earth's biomass? The biomass has on the order of 10^{45} carbon atoms. A reasonable estimate of the number of carbon atoms in a single replicating nanobot is about 10^{6}. (Note that this analysis is not very sensitive to the accuracy of these figures, only to the approximate order of magnitude.) This malevolent nanobot would need to create on the order of 10^{39} copies of itself to replace the biomass, which could be accomplished with 130 replications (each of which would potentially double the destroyed biomass). Rob Freitas has estimated a minimum replication time of approximately 100 seconds, so 130 replication cycles would require three to four hours.[3] However, the actual rate of destruction would be slower because biomass is not 'efficiently' laid out. The limiting factor would be the actual movement of the front of destruction. Nanobots cannot travel very quickly because of their small size. It's likely to take weeks for such a destructive process to circle the globe.

Based on this observation we can envision a more insidious possibility. In a two-phased attack, the nanobots take several weeks to spread throughout the biomass but use up an insignificant portion of the carbon atoms, say one out of every 10^{15}. At this extremely low level of concentration, the nanobots would be as stealthy as possible. Then, at an 'optimal' point, the second phase would begin with the seed nanobots expanding rapidly in place to destroy the biomass. For each seed nanobot to multiply itself a thousand trillionfold would require only about 50 binary replications, or about 90 minutes. With the nanobots having already spread out in position throughout the biomass, movement of the destructive wave front would no longer be a limiting factor.

The point is that without defenses, the available biomass could be destroyed by gray goo very rapidly. Clearly, we will need a nanotechnology immune system[4] in place *before* these scenarios become a possibility. This immune system would have to be capable of contending not just with obvious destruction but any potentially dangerous (stealthy) replication, even at very low concentration.

Eric Drexler, Robert Freitas, Ralph Merkle, Mike Treder, Chris Phoenix, and others have pointed out that future nanotech manufacturing devices can be created with safeguards that would prevent the accidental creation of self-replicating nanodevices.[5] However, this observation, although important, does not eliminate the threat of gray goo as I pointed out above. There are other reasons (beyond manufacturing) why self-replicating nanobots will need to be created. The nanotechnology immune system mentioned above, for example, will ultimately require self-replication, otherwise it would be unable to defend us against the

[3] http://www.kurzweilai.net/articles/art0142.html.

[4] More fully discussed in my book, *The Singularity is Near,* Chapter 8.

[5] http://www.crnano.org/BD-Goo.htm.

development of increasingly sophisticated types of goo. It is also likely to find extensive military applications. Moreover, a determined adversary or terrorist can defeat safeguards against unwanted self-replication; hence, the need for defense.

Bill Joy and other observers have pointed out that such an immune system would itself be a danger because of the potential of 'autoimmune' reactions (that is, the immune-system nanobots attacking the world they are supposed to defend). However, this possibility is not a compelling reason to avoid the creation of an immune system. No one would argue that humans would be better off without an immune system because of the potential of developing autoimmune diseases. Although the biological immune system can itself present a danger, humans would not last more than a few weeks (barring extraordinary efforts at isolation) without one. Furthermore, the development of a technological immune system for nanotechnology will happen even without explicit efforts to create one. This has effectively happened with regard to software viruses, creating an immune system not through a formal grand-design project but rather through incremental responses to each new challenge and by developing heuristic algorithms for early detection. We can expect the same thing will happen as challenges from nanotechnology-based dangers emerge. The point for public policy will be to specifically invest in these defensive technologies.

As a test case, we can take a small measure of comfort from how we have dealt with one recent technological challenge. There exists today a new fully nonbiological self-replicating entity that didn't exist just a few decades ago: the computer virus. When this form of destructive intruder first appeared, strong concerns were voiced that as they became more sophisticated, software pathogens had the potential to destroy the computer-network medium in which they live. Yet the 'immune system' that has evolved in response to this challenge has been largely effective. Although destructive self-replicating software entities do cause damage from time to time, the injury is but a small fraction of the benefit we receive from the computers and communication links that harbor them.

One might counter that computer viruses do not have the lethal potential of biological viruses or of destructive nanotechnology. This is not always the case; we rely on software to operate our 911 call centers, monitor patients in critical-care units, fly and land airplanes, guide intelligent weapons in our military campaigns, handle our financial transactions, operate our municipal utilities, and many other mission-critical tasks. To the extent that software viruses do not yet pose a lethal danger, however, this observation only strengthens my argument. The fact that computer viruses are not usually deadly to humans only means that more people are willing to create and release them. The vast majority of software virus authors would not release viruses if they thought they would kill people. It also means that our response to the danger is that much less intense. Conversely, when it comes to self-replicating entities that are potentially lethal on a large scale, our response on all levels will be vastly more serious.

Although software pathogens remain a concern, the danger exists today mostly at a nuisance level. Keep in mind that our success in combating them has taken place in an industry in which there is no regulation and minimal certification for practitioners. The largely unregulated computer industry is also enormously productive. One could argue that it has contributed more to our technological and economic progress than any other enterprise in human history.

But the battle concerning software viruses and the panoply of software pathogens will never end. We are becoming increasingly reliant on mission-critical software systems, and the sophistication and potential destructiveness of self-replicating software weapons will continue to escalate. When we have software running in our brains and bodies and controlling the world's nanobot immune system, the stakes will be immeasurably greater.

The right level of relinquishment

The only conceivable way that the accelerating pace of GNR technology advancement could be stopped would be through a worldwide totalitarian system that relinquishes the very idea of progress. Even this specter would be likely to fail in averting the dangers of GNR because the resulting underground activity would tend to favor the more destructive applications. This is because the responsible practitioners that we rely on to quickly develop defensive technologies would not have easy access to the needed tools. Fortunately, such a totalitarian outcome is unlikely because the increasing decentralization of knowledge is inherently a democratizing force.

I do think that relinquishment at the right level needs to be part of our ethical response to the dangers of 21st century technologies. One constructive example of this is the ethical guideline proposed by the Foresight Institute: namely, that nanotechnologists agree to relinquish the development of physical entities that can self-replicate in a natural environment. In my view, there are two exceptions to this guideline. First, we will ultimately need to provide a nanotechnology-based planetary immune system (nanobots embedded in the natural environment to protect against rogue self-replicating nanobots). Robert Freitas and I have discussed whether or not such an immune system would itself need to be self-replicating. Freitas writes: "A comprehensive surveillance system coupled with prepositioned resources—resources including high-capacity nonreplicating nanofactories able to churn our large numbers of nonreplicating defenders in response to specific threats—should suffice."[6] I agree with Freitas that a prepositioned immune system with the ability to augment the defenders will be sufficient in early stages. But once strong AI is merged with nanotechnology, and the ecology of nanoengineered entities becomes highly varied and complex, my own expectation is that we will find that the defending nanorobots need the ability to replicate in place quickly. Biological evolution essentially made the same 'discovery'. The other exception is the need for self-replicating nanobot-based probes to explore planetary systems outside of our solar system.

Broad relinquishment of GNR technologies would be unwise for several reasons. However, I do think we need to take seriously the increasingly strident voices that advocate it, even though many of these advocates are motivated by a general distrust of technology, and their proposals are not well-considered. Although blanket relinquishment is not the answer, rational fear could lead to irrational solutions, and those solutions may have severe negative consequences.

A summary of an overall strategy for defending against the downsides of emerging GNR technologies would include the following:

- We need to streamline the regulatory process for genetic and medical technologies. The

[6] Personal communication.

regulations do not impede the malevolent use of technology but significantly delay the needed defenses. As mentioned, we need to better balance the risks of new technology (for example, new medications) against the known harm of delay.

- A global program of confidential, random serum monitoring for unknown or evolving biological pathogens should be funded. Diagnostic tools exist to rapidly identify the existence of unknown protein or nucleic acid sequences. Intelligence is key to defense, and such a program could provide invaluable early warning of an impending epidemic. Such a 'pathogen sentinel' program has been proposed for many years by public health authorities but has never received adequate funding.
- Well-defined and targeted temporary moratoriums, such as the one that occurred in the genetics field in 1975, may be needed from time to time. But such moratoriums are unlikely to be necessary with nanotechnology. Broad efforts at relinquishing major areas of technology serve only to continue vast human suffering by delaying the beneficial aspects of new technologies, and actually make the dangers worse.
- Efforts to define safety and ethical guidelines for nanotechnology should continue. Such guidelines will inevitably become more detailed and refined as we get closer to molecular manufacturing.
- To create the political support to fund the efforts suggested above, it is necessary to *raise public awareness of these dangers*. Because, of course, there exists the downside of raising alarm and generating uninformed backing for broad antitechnology mandates, we also need to create a public understanding of the profound benefits of continuing advances in technology.
- These risks cut across international boundaries—which is, of course, nothing new; biological viruses, software viruses, and missiles already cross such boundaries with impunity. *International cooperation* was vital to containing the SARS virus and will become increasingly vital in confronting future challenges. Worldwide organizations such as the World Health Organization, which helped coordinate the SARS response, and is now dealing with the possibility of a bird flu pandemic, need to be strengthened.
- A contentious contemporary political issue is the need for preemptive action to combat threats, such as terrorists with access to weapons of mass destruction or rogue nations that support such terrorists. Such measures will always be controversial, but the potential need for them is clear. A nuclear explosion can destroy a city in seconds. A self-replicating pathogen, whether biological or nanotechnology-based, could destroy our civilization in a matter of days or weeks. We cannot always afford to wait for the massing of armies or other overt indications of ill intent before taking protective action.
- Intelligence agencies and policing authorities will have a vital role in forestalling the vast majority of potentially dangerous incidents. Their efforts need to involve the most powerful technologies available. For example, before this decade is out devices the size of dust particles will be able to carry out reconnaissance missions. When we reach the 2020s and have software running in our bodies and brains, government authorities will have a legitimate need on occasion to monitor these software streams. The potential for abuse of such powers is obvious. We will need to achieve a middle road of preventing catastrophic events while preserving our privacy and liberty.

The above approaches will be inadequate to deal with the danger from pathological R (strong AI). Our primary strategy in this area should be to optimize the likelihood that future nonbiological intelligence will reflect our values of liberty, tolerance, and respect for knowledge and diversity. The best way to accomplish this is to foster those values in our society today and going forward. If this sounds vague, it is. But there is no purely technical strategy that is workable in this area because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence. The nonbiological intelligence we are creating is and will be embedded in our societies and will reflect our values, as inconsistent and conflicting as these may appear to be. The transbiological phase will involve nonbiological intelligence deeply integrated with biological intelligence. This will amplify our abilities, and our application of these greater intellectual powers will be governed by the values of its creators. The transbiological era will ultimately give way to the postbiological era, but it is to be hoped that our values will remain influential. This strategy is certainly not foolproof, but it is the primary means we have today to influence the future course of strong AI.

Technology will remain a double-edged sword. It represents vast power to be used for all humankind's purposes. GNR will provide the means to overcome age-old problems such as illness and poverty, but will also empower destructive ideologies. We have no choice but to strengthen our defenses while we apply these quickening technologies to advance our human values, despite an apparent lack of consensus on what those values should be.

About the author:

Ray was the principal developer of the first omni-font OCR, the first print-to-speech reading machine, the first CCD flat-bed scanner, the first text-to-speech synthesizer, the first music synthesizer capable of recreating orchestral instruments, and the first commercially marketed large-vocabulary speech recognition.

Among Ray's many honors, he is the recipient of the MIT Lemelson Prize, the world's largest for innovation, and the National Medal of Technology, the nation's highest honor in technology, and has been inducted into the National Inventor's Hall of Fame. He has received twelve honorary Doctorates and honors from three U.S. presidents.

Ray has written five books, four of which have been national best sellers. His most recent book is *The Singularity is Near* (www.Singularity.com).