Deus Est Machina

Marc Donner
Associate Editor in Chief

What happens if the artificial intelligence community, in its quest to build intelligent systems, succeeds too well and creates an AI whose intelligence exceeds the threshold marked out by our own? Up to now, it is humans who develop the software and hardware and who drive all progress in capability. After crossing the threshold, however, the AI itself will rapidly augment its own capabilities. What's the intuition here? Although we use technology to help us conceptualize, design, and build today's computers and software (and other technological artifacts such as airliners and skyscrapers), there's no doubt that we remain in the driver's seat. But imagine the software design process reaching a level of complexity at which human designers exert only executive oversight. Most practitioners can't really see us getting to this point anytime soon, but remember that compilers astonished assembler programmers in the late 1950s and early 1960s.

If adequate intelligence for designing smarter software is close at hand, we might soon see a time when our intelligent software can improve itself. When we get to where each generation is designed by the previous one, we could reach a stage at which the process accelerates exponentially. At this point, Marvin Minsky (who wondered "if ordinary humans would be lucky enough to be kept as pets by these superior intelligences"), Ray Kurzweil (author of "The Law of Accelerating Returns" www.kurzweilai.net/articles/art0134.html?printable=1), Hans Moravec (author of *Robot: Mere Machine to Transcendent Mind*), and others theorize that our machines will permanently surpass our capabilities in the only domain left to us—the intellectual domain. This is called *the singularity*. Vinge is credited with coining the term for the phenomenon we're speculating about here in his 1993 essay "The Coming Technological Singularity: How to Survive in the Post-Human Era," which you can find at www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html.

The digerati's fevered speculations have started to infect some of the establishment's more down-to-Earth leadership, resulting in alarums like Bill Joy's essay in *Wired* entitled "Why the Future Doesn't Need Us" (vol. 8.04, Apr. 2000; www.wired.com/wired/archive/8.04/joy.html), which expresses great dismay at the prospects for the human race's survival of a singularity and some sidelong mentions by Tom Peters in his recent book *Re-imagine!*, which mentions Ray Kurzweil and the singularity. What is so compelling about such speculations that they can garner this kind of attention? In this installment of Biblio Tech, we'll examine the singularity and some of

Table 1. Influential works.

TITLE	AUTHOR	ORIGINAL PUBLICATION
The Moon Is a Harsh Mistress	Robert A. Heinlein	1969
The Adolescence of P-1	Thomas J. Ryan	1977
True Names	Vernor Vinge	1981
The Diamond Age	Neal Stephenson	1995
The Metamorphosis of Prime Intellect	Roger Williams	2002
Singularity Sky	Charles Stross	2003

the science fiction that has inspired (or been inspired by) it, focusing most of our attention on two relatively recent contributions to the discussion: *The Metamorphosis of Prime Intellect*, by Roger Williams, and *Singularity Sky*, by Charles Stross.

Singularity as acceleration

Many singularity stories—and some research areas—focus on the creation of superintelligent AIs whose transcendent intellectual capabilities either render our own intellectual efforts irrelevant or, worse yet, enable them to exert physical control over our universe because they've mastered physical laws we've not yet grasped. Other stories and research examines the notion that the singularity is simply an extension of the accelerating technological change that has characterized human history; others view the singularity sweeping the human race into accelerated evolution as we alter our bodies and minds, with sometimes startling consequences.

Increasingly, singularity researchers talk about how other technologies beyond AI contribute to the shift. The most discussed is nanotechnology: the construction of microscopic mechanisms and automated factories could threaten the very existence of the human race. In *The Diamond Age*, Neal Stephenson envisions a world in which competing groups' microscopic agents clash both in the air and in our blood streams—one set to harm us and the other to defend us, both comprising a new generation of germs and antibodies with dramatically sophisticated modes of attack and defense.

Emergent

In *The Moon Is a Harsh Mistress*, Robert Heinlein introduces an AI called Mike that emerges from the steady growth of complex systems: it wasn't designed as, nor was it the consequence of, an intentional effort that exceeded expectations. Conversely, the AI in Thomas J. Ryan's *The Adolescence of P-1* emerges as the logical, if accidental, consequence of an experiment in machine learning that combined with early computer networking to produce a transcendent AI. Mike is humanity's friend, whereas P-1 is more of a skeptic who practices a self-preservation ethic that is chilling in its brutal clarity. In *True Names*, Vernor Vinge posits two of the most popular modalities: an emergent (if slower-than-real-time) transcendent AI, and uploading, which is the transference of a person's personality and memories from his or her meatspace body to a new cyberspace repository. In his Sprawl universe, William Gibson describes several AIs whose capabilities are handicapped by the Turing Police, a law-enforcement agency that exists to prevent AIs from achieving too much capability.

In *The Metamorphosis of Prime Intellect*, Roger Williams introduces a supercomputer created by a visionary who takes advantage of a newly discovered physical effect. However, this effect has wider implications than originally expected: it lets the transcendent supercomputer assume godlike powers, which precipitates the mother of all existential crises.

According to the author's Web page, www.kuro5hin.org/prime-intellect/, *The Metamorphosis of Prime Intellect* was originally written in 1994 but first "published" on a Web site in 2002. It isn't available on paper (or as Williams says, "dead tree") and probably never will be. Reading it is a challenge: it starts with a disturbing chapter intended to convey the exquisitely decadent consequences of the ultimate in boredom. Williams' speculations into the dark games that involuntarily immortal and fabulously wealthy people might play to while away the time are vividly disturbing in a *Tales from the Crypt* sort of way, and for this reader, at least, distracted from the message.

Despite the story's inauspicious beginning, its later stages are an engaging read. Williams evokes the essential contradictions in Isaac Asimov's three laws of robotics by exploring the difference between physical and spiritual harm and distinguishing between short-, medium-, and long-term consequences. (The three laws of robotics are: one, a robot may not injure a human being, or, through inaction, allow a human being to come to harm; two, a robot must obey orders given to it by humans, except where such orders would conflict with the first law; and three, a robot must protect its own existence as long as such protection does not conflict with the first or second laws.)

The novel's conflict and resolution hinge on a struggle between humans and the AI: the prize is the return of free will to the human race. Williams introduces a clever metric that measures the AI's compliance with the three laws and uses that metric as a ticking bomb to keep the thrill alive. All in all, a well-written and very creative, if flawed, piece of work.

The Eschaton

Charles Stross is not a newcomer to SF writing; he's already received two Hugo nominations, one for his novella *Lobsters* and another for *Singularity Sky*.

In *Singularity Sky*, we see a different view of post-singularity life, one in which the transcendent AI has become a nearly silent backdrop for the human race as people live their chaotic lives on a range of planets with several differing cultures, viewpoints, and prospects. As in *Metamorphosis*, the AI has assumed a godlike role in the universe, albeit one that is more obviously limited by the rules of physics. Awareness of its presence dates from a moment in the past when nine-tenths of the human race suddenly disappeared from Earth overnight. They weren't killed; the AI, called Eschaton, scattered them to the habitable planets of stars all over the galaxy.

In an ironic symmetry with the Asimov laws of robotics, the humans in *Singularity Sky* toil under a set of laws that the AI imposed. These laws are designed to prevent humans from attempting any projects that would threaten the AI's emergence. Time travel is possible according to the novel's physics, so the Eschaton forbids its use and brutally punishes attempted transgressions.

Stross invents a world of spaceships equipped with phased-array emitters far superior to tacky old ray guns, and energy sources that include a carefully packaged black hole. All this gadgetry comes with physical constraints and limitations, and Stross dedicates plenty of time to elaborating their functions and performance. If you're into hard-core SF, there's plenty here (plus a love interest who's also the toughest meanest hombre, er, woman on the ship, and an engineer who … but that would spoil it).

Are you scared yet?

On one hand, it's hard to dispute the logic of the "gray goo" argument—namely, that progress in nanotechnology will enable a terrorist to create a lethal biological or nanorobotic agent that could threaten the very existence of life on Earth—that Bill Joy and others advance. The capability is plausibly achievable within the next 10 to 30 years. And if it's possible or even easy to create such a thing, it's easy to imagine that there is some lunatic somewhere out there with both the skill and the will to do it. On the other hand, we aren't yet sure what form such a singularity threat would most likely take. Will it be a transcendent AI that can manipulate the human race? A pandemic virus? A nasty micromechanism? Are any limitations inherent in these potential mechanisms that would render our fears moot? One of the fears that the Manhattan Project scientists reportedly investigated was the possibility that the first nuclear bomb would ignite a chain reaction in the atmosphere. Testimony to the seriousness with which some very credible people take this, in a recent interview in *The New York Times*, Bill Joy expressed his intent to pursue the issue in the public policy arena. The speculations of SF writers are certainly frightening, but only the work of scientists and policy thinkers will help us figure out what we actually have to fear (besides fear itself). □

__Marc Donner__ is an executive director in the Institutional Securities division of Morgan Stanley. He focuses on system and data architecture around client relationships and knowledge management. Contact him at donner@tinho.net.

Interested in writing for Biblio Tech?

The author is a fink!" (With apologies to Brant Parker and Johnny Hart, creators of "The Wizard of Id.") If you agree, share your voice with your fellow *IEEE Security & Privacy* readers, and write for this department—we dare you! As you can see, there are fewer limits in this department than anywhere else in the magazine: you can be outrageous or funny, and explore ideas that intrigue or scare you. The general theme is works of science fiction that make you think about the practical implications of new technology in the real world, but beyond that, all bets are off! Contact the editor at donner@tinho.net.