



# THE NEW YORKER

## ELEMENTS

Science, technology, and the things that make up our world.



« The Psychology of Online Comments | Main

OCTOBER 24, 2013

## WHY WE SHOULD THINK ABOUT THE THREAT OF ARTIFICIAL INTELLIGENCE

POSTED BY GARY MARCUS

If the New York *Times's* latest article is to be believed, artificial intelligence is moving so fast it sometimes seems almost “magical.” Self-driving cars have arrived; Siri can listen to your voice and find the nearest movie theatre; and I.B.M. just set the “Jeopardy”-conquering Watson to work on medicine, initially training medical students, perhaps eventually helping in diagnosis. Scarcely a month goes by without the announcement of a new A.I. product or technique. Yet, some of the enthusiasm may be premature: as I’ve noted previously, we still haven’t produced machines with common sense, vision, natural language processing, or the ability to create other machines. Our efforts at directly simulating human brains remain primitive.



Still, at some level, the only real difference between enthusiasts and skeptics is a time frame. The futurist and inventor Ray Kurzweil thinks true, human-level A.I. will be here in less than two decades. My estimate is at least double that, especially given how little progress has been made in computing common sense; the challenges in building A.I., especially at the software level, are much harder than Kurzweil lets on.

But a century from now, nobody will much care about how *long* it took, only what happened next. It's likely that machines will be smarter than us before the end of the century—not just at chess or trivia questions but at just about everything, from mathematics and engineering to science and medicine. There might be a few jobs left for entertainers, writers, and other creative types, but computers will eventually be able to program themselves, absorb vast quantities of new information, and reason in ways that we carbon-based units can only dimly imagine. And they will be able to do it every second of every day, without sleep or coffee breaks.

For some people, that future is a wonderful thing. Kurzweil has written about a rapturous singularity in which we merge with machines and upload our souls for immortality; Peter Diamandis has argued that advances in A.I. will be one key to ushering in a new era of "abundance," with enough food, water, and consumer gadgets for all. Skeptics like Eric Brynjolfsson and I have worried about the consequences of A.I. and robotics for employment. But even if you put aside the sort of worries about what super-advanced A.I. might do to the labor market, there's another concern, too: that powerful A.I. might threaten us more directly, by battling us for resources.

Most people see that sort of fear as silly science-fiction drivel—the stuff of “The Terminator” and “The Matrix.” To the extent that we plan for our medium-term future, we worry about asteroids, the decline of fossil fuels, and global warming, not robots. But a dark new book by James Barrat, “Our Final Invention: Artificial Intelligence and the End of the Human Era,” lays out a strong case for why we should be at least a little worried.

Barrat's core argument, which he borrows from the A.I. researcher Steve Omohundro, is that the drive for self-preservation and resource acquisition may be inherent in all goal-driven systems of a certain degree of intelligence. In Omohundro's words, “if it is smart enough, a robot that is designed to play chess might also want to be build a spaceship,” in order to obtain more resources for whatever goals it might have. A purely rational artificial intelligence, Barrat writes, might expand “its idea of self-preservation ... to include proactive attacks on future threats,” including, presumably, people who might be loathe to surrender their resources to the machine. Barrat worries that “without meticulous, countervailing instructions, a self-aware, self-improving, goal-seeking system will go to lengths we'd deem ridiculous to fulfill its goals,” even, perhaps, commandeering all the world's energy in order to maximize whatever calculation it happened to be interested in.

Of course, one could try to ban super-intelligent computers altogether. But “the competitive advantage—economic, military, even artistic—of every advance in automation is so compelling,” Vernor Vinge, the mathematician and science-fiction author, wrote, “that passing laws, or having customs, that forbid such things merely assures that someone else will.”

If machines will eventually overtake us, as virtually everyone in the A.I. field believes, the real question is about *values*: how we instill them in machines, and how we then negotiate with those machines if and when their values are likely to differ greatly from our own. As the Oxford philosopher Nick Bostrom argued:

We cannot blithely assume that a superintelligence will necessarily share any of the final values stereotypically associated with wisdom and intellectual development in humans—scientific curiosity, benevolent concern for others, spiritual enlightenment and contemplation, renunciation of material acquisitiveness, a taste for refined culture or for the simple pleasures in life, humility and selflessness, and so forth. It might be possible through deliberate effort to construct a superintelligence that values

such things, or to build one that values human welfare, moral goodness, or any other complex purpose that its designers might want it to serve. But it is no less possible—and probably technically easier—to build a superintelligence that places final value on nothing but calculating the decimals of pi.

The British cyberneticist Kevin Warwick once asked, “How can you reason, how can you bargain, how can you understand how that machine is thinking when it’s thinking in dimensions you can’t conceive of?”

If there is a hole in Barrat’s dark argument, it is in his glib presumption that if a robot is smart enough to play chess, it might also “want to build a spaceship”—and that tendencies toward self-preservation and resource acquisition are inherent in any sufficiently complex, goal-driven system. For now, most of the machines that are good enough to play chess, like I.B.M.’s Deep Blue, haven’t shown the slightest interest in acquiring resources.

But before we get complacent and decide there is nothing to worry about after all, it is important to realize that the goals of machines could change as they get smarter. Once computers can effectively reprogram themselves, and successively improve themselves, leading to a so-called “technological singularity” or “intelligence explosion,” the risks of machines outwitting humans in battles for resources and self-preservation cannot simply be dismissed.

One of the most pointed quotes in Barrat’s book belongs to the legendary serial A.I. entrepreneur Danny Hillis, who likens the upcoming shift to one of the greatest transitions in the history of biological evolution: “We’re at that point analogous to when single-celled organisms were turning into multi-celled organisms. We are amoeba and we can’t figure out what the hell this thing is that we’re creating.”

Already, advances in A.I. have created risks that we never dreamt of. With the advent of the Internet age and its Big Data explosion, “large amounts of data is being collected about us and then being fed to algorithms to make predictions,” Vaibhav Garg, a computer-risk specialist at Drexel University, told me. “We do not have the ability to know when the data is being collected, ensure that the data collected is correct, update the information, or provide the necessary context.” Few people would have even dreamt of this risk even twenty years ago. What risks lie ahead? Nobody really knows, but Barrat is right to ask.

*Photograph by John Vink/Magnum.*



Before long, artificial intelligence will stop looking to humans for upgrades and start seeking improvements on their own. (© Warner Brothers/Courtesy of Everett Collection)

## What Happens When Artificial Intelligence Turns On Us?

**In a new book, James Barrat warns that artificial intelligence will one day outsmart humans, and there is no guarantee that it will be benevolent**

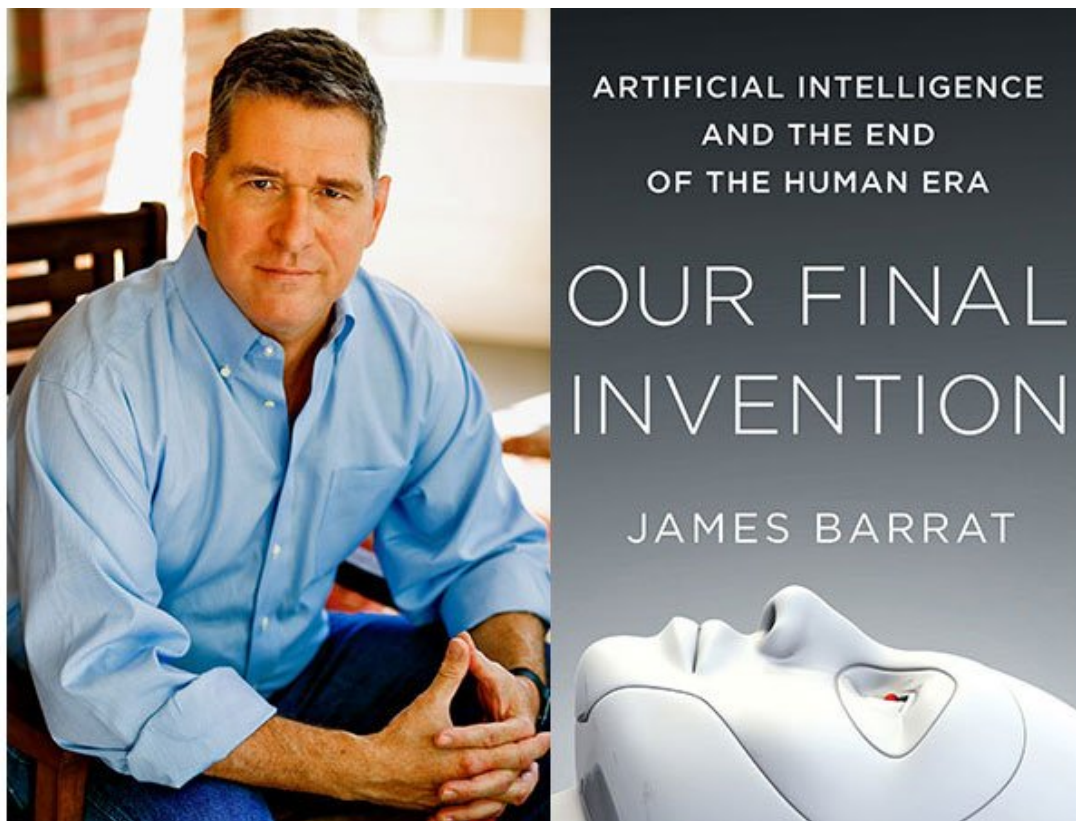
By [Erica R. Hendry](#)  
smithsonianmag.com  
January 21, 2014 6:23PM

Artificial intelligence has come a long way since R2-D2. These days, most millennials would be lost without smart GPS systems. Robots are already navigating battlefields, and drones may soon be delivering Amazon packages to our doorsteps.

Siri can solve complicated equations and tell you how to cook rice. She has even proven she can even respond to questions with a sense of humor.

But all of these advances depend on a user giving the A.I. direction. What would happen if GPS units decided they didn't want to

go to the dry cleaners, or worse, Siri decided she could become smarter without you around?



"Before we share the planet with super-intelligent machines, we must develop a science for understanding them. Otherwise, they'll take control," author James Barrat says of his new book, *Our Final Invention: Artificial Intelligence and the End of the Human Era*. (Courtesy of James Barrat)

These are just the tamest of outcomes James Barrat, an author and documentary filmmaker, forecasts in his new book, *Our Final Invention: Artificial Intelligence and the End of the Human Era*.

Before long, Barrat says, artificial intelligence—from Siri to drones and data mining systems—will stop looking to humans for upgrades and start seeking improvements on their own. And unlike the R2-D2s and HALs of science fiction, the A.I. of our future won't necessarily be friendly, he says: they could actually be what destroy us.

### **In a nutshell, can you explain your big idea?**

In this century, scientists will create machines with intelligence that equals and then surpasses our own. But before we share the planet with super-intelligent machines, we must develop a science for understanding them. Otherwise, they'll take control. And no, this isn't science fiction.

Scientists have already created machines that are better than humans at chess, *Jeopardy!*, navigation, data mining, search, theorem proving and countless other tasks. Eventually, machines will be created that are better than humans at A.I. research

At that point, they will be able to improve their own capabilities very quickly. These self-improving machines will pursue the goals they're created with, whether they be space exploration, playing chess or picking stocks. To succeed they'll seek and expend resources, be it energy or money. They'll seek to avoid the failure modes, like being switched off or unplugged. In short, they'll develop drives, including self-protection and resource acquisition—drives much like our own. They won't hesitate to beg, borrow, steal and worse to get what they need.

### **How did you get interested in this topic?**

I'm a documentary filmmaker. In 2000, I interviewed inventor Ray Kurzweil, roboticist Rodney Brooks and sci-fi legend Arthur C. Clarke for a TLC film about the making of the novel and film, *2001: A Space Odyssey*. The interviews explored the idea of the Hal 9000, and malevolent computers. Kurzweil's books have portrayed the A.I. future as a rapturous "singularity," a period in which technological advances outpace humans' ability to understand them. Yet he anticipated only good things emerging from A.I. that is strong enough to match and then surpass human intelligence. He predicts that we'll be able to reprogram the cells of

our bodies to defeat disease and aging. We'll develop super endurance with nanobots that deliver more oxygen than red blood cells. We'll supercharge our brains with computer implants so that we'll become superintelligent. And we'll port our brains to a more durable medium than our present "wetware" and live forever if we want to. Brooks was optimistic, insisting that A.I.-enhanced robots would be allies, not threats.

Scientist-turned-author Clarke, on the other hand, was pessimistic. He told me intelligence will win out, and humans would likely compete for survival with super-intelligent machines. He wasn't specific about what would happen when we share the planet with super-intelligent machines, but he felt it'd be a struggle for mankind that we wouldn't win.

That went against everything I had thought about A.I., so I began interviewing artificial intelligence experts.

### **What evidence do you have to support your idea?**

Advanced artificial intelligence is a dual-use technology, like nuclear fission, capable of great good or great harm. We're just starting to see the harm.

The NSA privacy scandal came about because the NSA developed very sophisticated data-mining tools. The agency used its power to plumb the metadata of millions of phone calls and the entirety of the Internet—critically, all email. Seduced by the power of data-mining A.I., an agency entrusted to protect the Constitution instead abused it. They developed tools too powerful for them to use responsibly.

Today, another ethical battle is brewing about making fully autonomous killer drones and battlefield robots powered by advanced A.I.—human-killers without humans in the loop. It's brewing between the Department of Defense and the drone and robot makers who are paid by the DOD, and people who think it's foolhardy and immoral to create intelligent killing machines. Those in favor of autonomous drones and battlefield robots argue that they'll be more moral—that is, less emotional, will target better and be more disciplined than human operators. Those against taking humans out of the loop are looking at drones' miserable history of killing civilians, and involvement in extralegal assassinations. Who shoulders the moral culpability when a robot kills? The robot makers, the robot users, or no one? Nevermind the technical hurdles of telling friend from foe.

In the longer term, as experts in my book argue, A.I. approaching human-level intelligence won't be easily controlled; unfortunately, super-intelligence doesn't imply benevolence. As A.I. theorist Eliezer Yudkowsky of MIRI [the Machine Intelligence Research Institute] puts it, "The A.I. does not love you, nor does it hate you, but you are made of atoms it can use for something else." If ethics can't be built into a machine, then we'll be creating super-intelligent psychopaths, creatures without moral compasses, and we won't be their masters for long.

### **What is new about your thinking?**

Individuals and groups as diverse as American computer scientist Bill Joy and MIRI have long warned that we have much to fear from machines whose intelligence eclipses our own. In *Our Final Invention*, I argue that A.I. will also be misused on the development path to human-level intelligence. Between today and the day when scientists create human-level intelligence, we'll have A.I.-related mistakes and criminal applications.

### **Why hasn't more been done, or, what is being done to stop AI from turning on us?**

There's not one reason, but many. Some experts don't believe we're close enough to creating human-level artificial intelligence and beyond to worry about its risks. Many A.I. makers win contracts with the Defense Advanced Research Projects Agency [DARPA] and don't want to raise issues they consider political. The normalcy bias is a cognitive bias that prevents people from reacting to disasters and disasters in the making—that's definitely part of it. But a lot of A.I. makers are doing something. Check out the scientists who advise MIRI. And, a lot more will get involved once the dangers of advanced A.I. enter mainstream dialogue.

### **Can you describe a moment when you knew this was big?**

We humans steer the future not because we're the fastest or the strongest creatures on the planet, but because we're the smartest. When we share the planet with creatures smarter than ourselves, they'll steer the future. When I understood this idea, I felt I was writing about the most important question of our time.

### **Every big thinker has predecessors whose work was crucial to his discovery. Who gave you the foundation to build your idea?**

The foundations of A.I. risk analysis were developed by mathematician I. J. Good, science fiction writer Vernor Vinge, and others

including A.I. developer [Steve Omohundro](#). Today, MIRI and Oxford's Future of Humanity Institute are almost alone in addressing this problem. *Our Final Invention* has about 30 pages of endnotes acknowledging these thinkers.

### **In researching and developing your idea, what has been the high point? And the low point?**

The high points were writing *Our Final Invention*, and my ongoing dialogue with A.I. makers and theorists. People who program A.I. are aware of the safety issues and want to help come up with safeguards. For instance, MIRI is working on creating "friendly" A.I.

Computer scientist and theorist Steve Omohundro has advocated a "scaffolding" approach, in which provably safe A.I. helps build the next generation of A.I. to ensure that it too is safe. Then that A.I. does the same, and so on. I think a public-private partnership has to be created to bring A.I.-makers together to share ideas about security—something like the International Atomic Energy Agency, but in partnership with corporations. The low points? Realizing that the best, most advanced A.I. technology will be used to create weapons. And those weapons eventually will turn against us.

### **What two or three people are most likely to try to refute your argument? Why?**

Inventor Ray Kurzweil is the chief apologist for advanced technologies. In my two interviews with him, he claimed that we would meld with the A.I. technologies through cognitive enhancements. Kurzweil and people broadly called transhumanists and singularitarians think A.I. and ultimately artificial general intelligence and beyond will evolve with us. For instance, computer implants will enhance our brains' speed and overall capabilities. Eventually, we'll develop the technology to transport our intelligence and consciousness into computers. Then super-intelligence will be at least partly human, which in theory would ensure super-intelligence was "safe."

For many reasons, I'm not a fan of this point of view. Trouble is, we humans aren't reliably safe, and it seems unlikely that super-intelligent humans will be either. We have no idea what happens to a human's ethics after their intelligence is boosted. We have a biological basis for aggression that machines lack. Super-intelligence could very well be an aggression multiplier.

### **Who will be most affected by this idea?**

Everyone on the planet has much to fear from the unregulated development of super-intelligent machines. An intelligence race is going on right now. Achieving A.G.I. is job number one for Google, IBM and many smaller companies like Vicarious and Deep Thought, as well as DARPA, the NSA and governments and companies abroad. Profit is the main motivation for that race. Imagine one likely goal: a virtual human brain at the price of a computer. It would be the most lucrative commodity in history. Imagine banks of thousands of PhD quality brains working 24/7 on pharmaceutical development, cancer research, weapons development and much more. Who wouldn't want to buy that technology?

Meanwhile, 56 nations are developing battlefield robots, and the drive is to make them, and drones, autonomous. They will be machines that kill, unsupervised by humans. Impoverished nations will be hurt most by autonomous drones and battlefield robots. Initially, only rich countries will be able to afford autonomous kill bots, so rich nations will wield these weapons against human soldiers from impoverished nations.

### **How might it change life, as we know it?**

Imagine: in as little as a decade, a half-dozen companies and nations field computers that rival or surpass human intelligence. Imagine what happens when those computers become expert at programming smart computers. Soon we'll be sharing the planet with machines thousands or millions of times more intelligent than we are. And, all the while, each generation of this technology will be weaponized. Unregulated, it will be catastrophic.

### **What questions are left unanswered?**

Solutions. The obvious solution would be to give the machines a moral sense that makes them value human life and property. But programming ethics into a machine turns out to be extremely hard. Moral norms differ from culture to culture, they change over time, and they're contextual. If we humans can't agree on when life begins, how can we tell a machine to protect life? Do we really want to be safe, or do we really want to be free? We can debate it all day and not reach a consensus, so how can we possibly program it?

We also, as I mentioned earlier, need to get A.I. developers together. In the 1970s, recombinant DNA researchers decided to suspend research and get together for a conference at Asilomar in Pacific Grove, California. They developed basic safety protocols like "don't track the DNA out on your shoes," for fear of contaminating the environment with genetic works in progress. Because of the "Asilomar Guidelines," the world benefits from genetically modified crops, and gene therapy looks

promising. So far as we know, accidents were avoided. It's time for an Asilomar Conference for A.I.

### **What's standing in the way?**

A huge economic wind propels the development of advanced A.I. Human-level intelligence at the price of a computer will be the hottest commodity in history. Google and IBM won't want to share their secrets with the public or competitors. The Department of Defense won't want to open their labs to China and Israel, and vice-versa. Public awareness has to push policy towards openness and public-private partnerships designed to ensure safety.

### **What is next for you?**

I'm a documentary filmmaker, so of course I'm thinking about a film version of *Our Final Invention*.

### **Tags**

[Artificial Intelligence](#) [Big Ideas](#) [Futurism](#)

### **About Erica R. Hendry**



Erica R. Hendry is the innovations reporter/producer for Smithsonian.com

[Read more from this author](#) | [Follow @ericarhendry](#)



[Return to the Article](#)

December 6, 2013

## Our Final Invention: How the Human Race Goes and Gets Itself Killed

By [Greg Scoblete](#)

We worry about robots.

Hardly a day goes by where we're not reminded about how robots are [taking our jobs](#) and hollowing out the middle class. The worry is so acute that economists are busy devising new social contracts to cope with a [potentially enormous class of obsolete humans](#).

Documentarian James Barrat, author of [Our Final Invention: Artificial Intelligence and the End of the Human Era](#), is worried about robots too. Only he's not worried about them taking our jobs. He's worried about them exterminating the human race.

I'll repeat that: In 267 brisk pages, Barrat lays out just how the artificial intelligence (AI) that companies like Google and governments like our own are racing to perfect could -- indeed, likely will -- advance to the point where it will literally destroy all human life on Earth. Not put it out of work. Not meld with it in a utopian fusion. Destroy it.

### Wait, What?

I'll grant you that this premise sounds a bit.... dramatic, the product of one too many *Terminator* screenings. But after approaching the topic with some skepticism, it became increasingly clear to me that Barrat has written an extremely important book with a thesis that is worrisomely plausible. It deserves to be read widely. And to be clear, Barrat's is not a lone voice -- the book is rife with interviews of numerous computer scientists and AI researchers who share his concerns about the potentially devastating consequences of advanced AI. There are even [think tanks](#) devoted to exploring and mitigating the risks. But to date, this worry has been obscure.

In Barrat's telling, we are on the brink of creating machines that will be as intelligent as humans. Specific timelines vary, but the broad-brush estimates place the emergence of human-level AI at between 2020 and 2050. This human-level AI (referred to as "artificial general intelligence" or AGI) is worrisome enough, seeing the damage human intelligence often produces, but it's what happens next that really concerns Barrat. That is, once we have achieved AGI, the AGI will go on to achieve something called artificial superintelligence (ASI) -- that is, an intelligence that exceeds -- vastly exceeds -- human-level intelligence.

Barrat devotes a substantial portion of the book explaining how AI will advance to AGI and how AGI inevitably leads to ASI. Much of it hinges on how we are developing AGI itself. To reach AGI, we are teaching machines to learn. The techniques vary -- some researchers approach it through something akin to the brute-force memorization of facts and images, others through a trial-and-error process that mimics genetic evolution, others by attempting to reverse engineer the human brain -- but the common thread stitching these efforts together is the creation of machines that constantly learn and then use this knowledge to improve themselves.

The implications of this are obvious. Once a machine built this way reaches human-level intelligence, it won't stop there. It will keep learning and improving. It will, Barrat claims, reach a point that other computer scientists have dubbed an "intelligence explosion" -- an onrushing feedback loop where an intelligence makes itself smarter thereby getting even better at making itself smarter. This is, to be sure, a theoretical concept, but it is one that many AI researchers see as plausible, if not inevitable. Through a relentless process of debugging and rewriting its code, our self-learning, self-programming AGI experiences a "hard take off" and rockets past what mere flesh and blood brains are capable of.

And here's where things get interesting. And by interesting I mean terrible.

---

### Goodbye, Humanity

When (and Barrat is emphatic that this is a matter of when, not if) humanity creates ASI it will have introduced into the world an intelligence greater than our own. This would be an existential event. Humanity has held pride of place on planet Earth because of our superior intelligence. In a world with ASI, we will no longer be the smartest game in town.

To Barrat, and other concerned researchers quoted in the book, this is a lethal predicament. At first, the relation between a human intellect and that of an ASI may be like that of an ape's to a human, but as ASI continues its process of perpetual self-improvement, the gulf widens.

At some point, the relation between ASI and human intelligence mirrors that of a human to an ant.

Needless to say, that's not a good place for humanity to be.

And here's the kicker. Barrat argues that the time it will take for ASI to surpass human level intelligence, rendering us ant-like in comparison, could be a matter of days, if not mere hours, after it is created. Worse (it keeps getting worse), human researchers may not even know they have created this potent ASI until it is too late to attempt to contain it. An ASI birthed in a supercomputer may choose, Barrat writes, to hide itself and its capabilities lest the human masters it knows so much about it, attempt to shut it down. Then, it would silently replicate itself and spread. With no need to eat and sleep and with an intelligence that is constantly improving and war-gaming survival strategies, ASI could hide, wait and grow its capabilities while humanity plods along, blissfully unaware.

Though we have played a role in creating it, the intelligence we would be faced with would be completely alien. It would not be a human's mind, with its experiences, emotions and logic, or lack thereof. We could not anticipate what ASI would do because we simply do not "think" like it would. In fact, we've already arrived at the alarming point where we do not understand what the machines we've created do. Barrat describes how the makers of Watson, IBM's Jeopardy winning supercomputer, could not understand how the computer was arriving at its correct answers. Its behavior was unpredictable to its creators -- and the mysterious Watson is not the only such inscrutable "black box" system in existence today, nor is it even a full-fledged AGI, let alone ASI.

Barrat grapples with two big questions in the book. The first is why an ASI necessarily leads to human extinction. Aren't we programming it? Why couldn't humanity leverage it, like we do any technology, to make our lives better? Wouldn't we program in safeguards to prevent an "intelligence explosion" or, at a minimum, contain one when it bursts?

According to Barrat, the answer is almost certainly no. Most of the major players in AI are barely concerned with safety, if at all. Even if they were, there are too many ways for AI to make an end-run around our safeguards (remember, these are human-safeguards matched up with an intelligence that will equal and then quickly exceed it). Programming "friendly AI" is also difficult, given that even the best computer code is rife with error and complex systems can suffer catastrophic failures that are entirely unforeseen by their creators. Barrat doesn't say the picture is utterly hopeless. It's possible, he writes, that with extremely careful planning humanity could contain a super-human intelligence -- but this is not the manner in which AI development is unfolding. It's being done by defense agencies around the world in the dark. It's being done by private companies who reveal very little about what it is they're doing. Since the financial and security benefits of a working AGI could be huge, there's very little incentive to pump the breaks before the more problematic ASI can emerge.

Moreover, ASI is unlikely to exterminate us in a bout of *Terminator*-esque malevolence, but simply as a byproduct of its very existence. Computers, like humans, need energy and in a competition for resources, ASI would no more seek to preserve our access to vital resources than we worry about where an ant's next meal will come from. We cannot assume ASI empathy, Barrat writes, nor can we assume that whatever moral strictures we program in will be adhered to. If we do achieve ASI, we will be in completely unknown territory. (But don't rule out a *Terminator* scenario altogether -- one of the biggest drivers of AI research is the Pentagon's DARPA and they are, quite explicitly, building killer robots. Presumably other well-funded defense labs, in China and Russia, are doing similar work as well.)

Barrat is particularly effective in rebutting devotees of the Singularity -- the techno-optimism popularized by futurist Ray Kurzweil (now at Google, a company investing millions in AI research). Kurzweil and his fellow Singularitins also believe that ASI is inevitable only they view it as a force that will liberate and transform humanity for the good, delivering the dream of immortality and solving all of our problems. Indeed, they agree with Barrat that the "intelligence explosion" signals the end of humanity as we know it, only they view this as a benign development with humanity and ASI merging in a "transhuman" fusion.

If this sounds suspiciously like an end-times cult that's because, in its crudest expression, it is (one that just happens to be filled with more than a few brilliant computer scientists and venture capitalists). Barrat forcefully contends that even its more nuanced formulation is an irredeemably optimistic interpretation of future trends and human nature. In fact, efforts to merge ASI with human bodies is even *more* likely to birth a catastrophe because of the malevolence that humanity is capable of.

The next question, and the one with the less satisfactory answer, is just *how* ASI would exterminate us. How does an algorithm, a piece of programming lying on a supercomputer, reach out into the "real" world and harm us? Barrat raises a few scenarios -- it could leverage future nano-technologies to strip us down at the molecular level, it could shut down our electrical grids and turn the electronic devices we rely on against us -- but doesn't do nearly as much dot-connecting between ASI as a piece of computer code and the physical mechanics of how this code will be instrumental in our demise as he does in establishing the probability of achieving ASI.

That's not to say the dots don't exist, though. Consider the world we live in right now. Malware can travel [through thin air](#). Our homes, cars, planes, hospitals, refrigerators, ovens (even our [forks](#) for God's sake) connect to an "internet of things" which is itself spreading on the backs of ubiquitous wireless broadband. We are steadily [integrating electronics](#) inside our bodies. And a few mistaken lines of code in the most dangerous computer virus ever created (Stuxnet) caused it to wiggle free of its initial target and [travel the world](#). Now extrapolate these trends out to 2040 and you realize that ASI will be born into a world that is utterly intertwined and dependent on the virtual, machine world -- and vulnerable to it. (Indeed one AI researcher Barrat interviews argues that this is precisely why we need to create ASI as fast as possible, while its ability to harm us is still relatively constrained.)

What we're left with is something beyond dystopia. Even in the bleakest sci-fi tales, a scrappy contingent of the human race is left to duke it out with their runaway machines. If *Our Final Invention* is correct, there will be no such heroics, just the remorseless evolutionary logic

that has seen so many other species wiped off the face of the Earth at the hands of a superior predator.

Indeed, it's telling that both AI-optimists like Kurzweil and pessimists like Barrat reach the same basic conclusion: humanity as we know it will not survive the birth of intelligent machines.

No wonder we're worried about robots.

*Greg Scoblete (@GregScoblete) is the editor of RealClearTechnology and an editor on RealClearWorld. He is the co-author of [From Fleeting to Forever: A Guide to Enjoying and Preserving Your Digital Photos and Videos](#).*

*(Image: St. Martin's Press)*

**Page Printed from:**

**[http://www.realcleartechology.com/articles/2013/12/06/our\\_final\\_invention\\_how\\_the\\_human\\_race\\_goes\\_and\\_gets\\_itself\\_killed\\_816-full.html](http://www.realcleartechology.com/articles/2013/12/06/our_final_invention_how_the_human_race_goes_and_gets_itself_killed_816-full.html)** at December 06, 2013 - 09:18:46 AM EST

## Artificial intelligence: Our final invention?

By [Matt Miller](#), Wednesday, December 18, 9:08 AM

Even when our debates seem petty, you can't say national politics doesn't deal with weighty matters, from jobs to inequality to affordable health care and more. But lately I've become obsessed with an issue so daunting it makes even the biggest "normal" questions of public life seem tiny. I'm talking about the risks posed by "runaway" artificial intelligence (AI). What happens when we share the planet with self-aware, self-improving machines that evolve beyond our ability to control or understand? Are we creating machines that are destined to destroy us?

I know when I put it this way it sounds like science fiction, or the ravings of a crank. So let me explain how I came to put this on your screen.

A few years ago I read chunks of Ray Kurzweil's book "[The Singularity Is Near](#)." Kurzweil argued that what sets our age apart from all previous ones is the *accelerating* pace of technological advance — an acceleration made possible by the digitization of everything. Because of this unprecedented pace of change, he said, we're just a few decades away from basically meshing with computers and transcending human biology (think Google, only much better, inside your head). This development will supercharge notions of "intelligence," Kurzweil predicted, and even make it possible to upload digitized versions of our brains to the cloud so that some form of "us" lives forever.

Mind-blowing and unsettling stuff, to say the least. If Kurzweil's right, I recall thinking, what should I tell my daughter about how to live — or even about what it means to be human?

Kurzweil has since become enshrined as America's uber-optimist on these trends. He and other evangelists say accelerating technology will soon equip us to solve our greatest energy, education, health and climate challenges en route to extending the human lifespan indefinitely.

But a camp of worrywarts has sprung up as well. The skeptics fear that a toxic mix of artificial intelligence, robotics and bio- and nanotechnology could make previous threats of nuclear devastation seem "easy" to manage by comparison. These people aren't cranks. They're folks like [Jaan Tallinn](#), the 41-year-old Estonian programming whiz who helped create Skype and now fears he's more likely to die from some AI advance run amok than from cancer or heart disease. Or Lord Martin Rees, a dean of Britain's science establishment whose last book bore the upbeat title, "[Our Final Century](#)" and who with Tallinn has launched the [Center for the Study of Existential Risk](#) at Cambridge to think through how bad things could get and what to do about it.

Now comes James Barrat with a new book — "[Our Final Invention: Artificial Intelligence and the End of the Human Era](#)" — that accessibly chronicles these risks and how a number of top AI researchers and observers see them. If you read just one book that makes you confront scary high-tech realities that we'll soon have no choice but to address, make it this one.

In an [interview the other day for my podcast show](#) “This...Is Interesting,” Barrat, an Annapolis-based documentary filmmaker, noted that every technology since fire has brought both promise and peril. How should we weigh the balance with AI?

It turns out that in talking with dozens in the field, Barrat found that everyone is aware of the potential risks of “runaway AI,” but no one spends any time on it. Why not? Barrat surmised that “normalcy bias” — which holds that if something awful hasn’t happened until now it probably won’t happen in the future — accounts for the silence.

Many AI researchers simply assume we’ll be able to build “friendly AI,” systems that are programmed with our values and with respect for humans as their creators. When pressed, however, most researchers admit to Barrat that this is wishful thinking.

The better question may be this: Once our machines become literally millions or trillions of times smarter than we are (in terms of processing power and the capabilities this enables), what reason is there to think they’ll view us any differently than we view ants or pets?

The military applications of AI guarantee a new arms race, which the Pentagon and the Chinese are already quietly engaged in. AI’s endless commercial applications assure an equally competitive sprint by major firms. IBM, Barrat said, has been laudably transparent with its plans to turn its Jeopardy-playing “Watson” into a peerless medical diagnostician. But Google — which hired Kurzweil earlier this year as director of engineering, and which also has a former head of the Pentagon’s advanced research agency on the payroll — isn’t talking.

Meanwhile, the military is already debating the ethical implications of giving autonomous drones the authority to use lethal force without human intervention. Barrat sees the coming AI crisis as analogous to nuclear fission and recombinant DNA, which inspired passionate debates over how to pursue these technologies responsibly.

This spring Hollywood will weigh in with “[Transcendence](#),” starring Johnny Depp as an AI researcher targeted by violent extremists who think we’re crossing a Rubicon that will be a disaster for humanity.

At the end of our interview, I asked Barrat what I thought was a joke. I know you’ve got a grim view of what may lie ahead, I said, but does that mean you’re buying property for your family on a desert island just in case?

“I don’t want to really scare you,” he said, after half a chuckle. “But it was alarming how many people I talked to who are highly placed people in AI who have retreats that are sort of ‘bug out’ houses” to which they could flee if it all hits the fan.

Whoa.

It’s time to take this conversation beyond a few hundred technology sector insiders or even those reached by Barrat’s indispensable wake-up call. In his State of the Union address next month, President Obama should set up a presidential commission on the promise and perils of artificial intelligence to kick-start the national debate AI’s future demands.

*Read more from [Matt Miller’s archive](#) or [follow him on Twitter](#).*

Read more on this topic: [The Post’s View: Congress needs to address the threat posed by 3-D printed guns](#)  
[Tom Jackman: When loved ones go missing, don’t count on technology to save them](#)  
[Evan Marwell: Using](#)

fiber optics to bring schools up to Internet speed



### **Map Your Flood Risk**

Find Floodplan Maps, Facts, FAQs, Your Flood Risk Profile and More!  
[www.floodsmart.gov](http://www.floodsmart.gov)

### **360 Savings Account**

360 Savings. No fees. No minimums. Nothing standing in your way. Learn More.  
[www.capitalone360.com](http://www.capitalone360.com)

### **Brain Training Games**

Challenge memory and attention with scientific brain games.  
[www.lumosity.com](http://www.lumosity.com)

**Buy a link here**

© The Washington Post Company

## When Robots Take Over, What Happens to Us?

PAUL WALDMAN

NOVEMBER 11, 2013

**We interviewed James Barrat, author of *Our Final Invention: Artificial Intelligence and the End of the Human Era*, to see what happens when we're no longer the most intelligent inhabitants of Earth.**

--

Artificial intelligence has a long way to go before computers are as intelligent as humans. But progress is happening rapidly, in everything from logical reasoning to facial and speech recognition. With steady improvements in memory, processing power, and programming, the question isn't *if* a computer will ever be as smart as a human, but only how long it will take. And once computers are as smart as people, they'll keep getting smarter, in short order become much, much smarter than people. When artificial intelligence (AI) becomes artificial superintelligence (ASI), the real problems begin.

In his new book *Our Final Invention: Artificial Intelligence and the End of the Human Era*, James Barrat argues that we need to begin thinking now about how artificial intelligences will treat their creators when they can think faster, reason better, and understand more than any human. These questions were long the province of thrilling (if not always realistic) science fiction, but Barrat warns that the consequences could indeed be catastrophic. I spoke with him about his book, the dangers of ASI, and whether we're all doomed.

**Your basic thesis is that even if we don't know exactly how long it will take, eventually artificial intelligence will surpass human intelligence, and once they're smarter than we are, we are in serious trouble. This is an idea people are familiar with; there are lots of sci-fi stories about homicidal AIs like HAL or Skynet. But you argue that it may be more likely that super-intelligent AI will be simply indifferent to the fate of humanity, and that could be just as dangerous for us. Can you explain?**

First, I think we've been inoculated to the threat of advanced AI by science fiction. We've had so much fun with Hollywood tropes like Terminator and of course the Hal 9000 that we don't take the threat seriously. But as **Bill Joy** once said, "Just because you saw it in a movie doesn't mean it can't happen."

Superintelligence in no way implies benevolence. Your laptop doesn't like you or dislike you anymore than your toaster does— why do we believe an intelligent machine will be different? We humans have a bad habit of



imputing motive to objects and phenomena—anthropomorphizing. If it's thundering outside the gods must be angry. We see friendly faces in clouds. We anticipate that because we create an artifact, like an intelligent machine, it will be grateful for its existence, and want to serve and protect us.

But these are our qualities, not machines'. Furthermore, at an advanced level, as I write in *Our Final Invention*, citing the work of AI-maker and theorist Steve Omohundro, artificial intelligence will have drives much like our own, including self-protection and resource acquisition. It will want to achieve its goals and marshal sufficient resources to do so. It will want to avoid being turned off. When its goals collide with ours it will have no basis for valuing our goals, and use whatever means are at its disposal for achieving its goals.

**The immediate answer many people would give to the threat is, "Well, just program them not to hurt us," with some kind of updated version of Isaac Asimov's *Three Laws of Robotics*. I'm guessing that's no easy task.**

That's right, it's extremely difficult. Asimov's Three Laws are often cited as a cure-all for controlling ASI. In fact they were created to generate tension and stories. His classic *I, Robot* is a catalogue of unintended consequences caused by conflicts among the three laws. Not only are our values hard to give to a machine, our values change from culture to culture, religion to religion, and over time. We can't agree on when life

begins, so how can we reach a consensus about the qualities of life we want to protect? And will those values make sense in 100 years?

ADVERTISEMENT

loading

```
<IFRAME ID="4EA03CC6B7F8A" NAME="4EA03CC6B7F8A" SRC="HTTP://OX-D.PROSPECT.ORG/W/1.0/AFR?AUID=98717&CB=INSERT_RANDOM_NUMBER_HERE" FRAMEBORDER="0" FRAMESPACING="0" SCROLLING="NO" WIDTH="300" HEIGHT="250"><A HREF="HTTP://OX-D.PROSPECT.ORG/W/1.0/RC?CS=4EA03CC6B7F8A&CB=INSERT_RANDOM_NUMBER_HERE"><IMG SRC="HTTP://OX-D.PROSPECT.ORG/W/1.0/AI?AUID=98717&CS=4EA03CC6B7F8A&CB=INSERT_RANDOM_NUMBER_HERE" BORDER="0" ALT=""></A></IFRAME>
```

**When you're discussing our efforts to contain an AI many times smarter than us, you make an analogy to waking up in a prison run by mice (with whom you can communicate). My takeaway from that was pretty depressing. Of course you'd be able to manipulate the mice into letting you go free, and it would probably be just as easy for an artificial superintelligence to get us to do what it wants. Does that mean any kind of technological means of containing it will inevitably fail?**

*Our Final Invention* is both a warning and a call for ideas about how to govern superintelligence. I think we'll struggle mortally with this problem, and there aren't a lot of solutions out there—I've been looking. Ray Kurzweil, who's portrait of the future is very rosy, concedes that superior intelligence won't be contained. His solution is to merge with it. The 1975 Asilomar Conference on Recombinant DNA is a good model of what should happen. Researchers suspended work and

got together to establish basic safety protocols, like "don't track the DNA out on your shoes." It worked, and now we're benefitting from gene therapy and better crops, with no horrendous accidents so far. MIRI (the Machine Intelligence Research Institute) advocates creating the first superintelligence with friendliness encoded, among other steps, but that's hard to do. Bottom line—before we share the planet with superintelligent machines we need a science for understanding and controlling them.

**But as you point out, it would be extremely difficult in practical terms to ban a particular kind of AI—if we don't build it, someone else will, and there will always be what seem to them like very good reasons to do so. With people all over the world working on these technologies, how can we impose any kind of stricture that will prevent the outcomes we're afraid of?**

Human-level intelligence at the price of a computer will be the most lucrative commodity in the history of the world. Imagine banks of thousands of PhD quality brains working on cancer research, climate modeling, weapons development. With those enticements, how do you get competing researchers and countries to the table to discuss safety? My answer is to write a book, make films, get people aware and involved, and start a private-public partnership targeted at safety. Government and industry have to get together. For that to happen, we must give people the resources they need to understand a problem that's going to deeply affect their lives. Public pressure is all we've got to get people

to the table. If we wait to be motivated by horrendous accidents and weaponization, as we have with nuclear fission, then we'll have waited too long.

**Beyond the threat of annihilation, one of the most disturbing parts of this vision is the idea that we'll eventually reach the point at which humans are no longer the most important actors on planet Earth. There's another species (if you will) with more capability and power to make the big decisions, and we're here at their indulgence, even if for the moment they're treating us humanely. If we're a secondary species, how do you think that will affect how we think about what it means to be human?**

That's right, we humans steer the future not because we're the fastest or strongest creatures, but because we're the smartest. When we share the planet with creatures smarter than we are, they'll steer the future. For a simile, look at how we treat intelligent animals - they're at Seaworld, they're bushmeat, they're in zoos, or they're endangered. Of course the Singularitarians believe that the superintelligence will be ours—we'll be transhuman. I'm deeply skeptical of that one-sided good news story.

**As you were writing this book, were there times you thought, "That's it. We're doomed. Nothing can be done"?**

Yes, and I thought it was curious to be alive and aware within the time window in which we might be able to change that future, a twist on the anthropic principal.

But having hope about seemingly hopeless odds is a moral choice. Perhaps we'll get wise to the dangers in time. Perhaps we'll learn after a survivable accident. Perhaps enough people will realize that advanced AI is a dual use technology, like nuclear fission. The world was introduced to fission at Hiroshima. Then we as a species spent the next 50 years with a gun pointed at our own heads. We can't survive that abrupt an introduction to superintelligence. And we need a better maintenance plan than fission's mutually assured destruction.

## ADVANCE PRAISE FOR OUR FINAL INVENTION

*A hard-hitting book about the most important topic of this century and possibly beyond — the issue of whether our species can survive. I wish it was science fiction but I know it's not.”*

Jaan Tallinn, co-founder of Skype

*The compelling story of humanity's most critical challenge. A Silent Spring for the 21st Century.”*

Michael Vassar, former President of the Singularity Institute

*Barrat's book is excellently written and deeply researched. It does a great job of communicating to general readers the danger of mistakes in AI design and implementation.”*

Bill Hibbard, author of Super-Intelligent Machines

*Our Final Invention is a thrilling detective story, and also the best book yet written on the most important problem of the 21st century.”*

Luke Muehlhauser, Executive Director of the Machine Intelligence Research Institute

*An important and disturbing book.”*

Huw Price, co-founder, Cambridge University Center for the Study of Existential Risk

## ***Our Final Invention: Is AI the Defining Issue for Humanity?***

By Seth Baum | October 11, 2013

Humanity today faces incredible threats and opportunities: climate change, nuclear weapons, biotechnology, nanotechnology, and much, much more. But some people argue that these things are all trumped by one: artificial intelligence (AI). To date, this argument has been confined mainly to science fiction and a small circle of scholars and enthusiasts. Enter documentarian [James Barrat](#), whose new book [Our Final Invention](#) states the case for (and against) AI in clear, plain language.

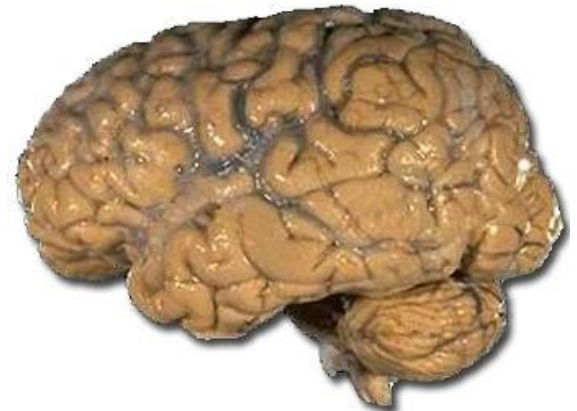
ADVERTISEMENT

*Disclosure: I know Barrat personally. He sent me a free advance copy in hope that I would write a review. The book also [cites research of mine](#). And I am an unpaid Research Advisor to the [Machine Intelligence Research Institute](#), which is discussed heavily in the book. But while I have some incentive to say nice things, I will not be sparing in what (modest) criticism I have.*

The central idea is haltingly simple. Intelligence could be the key trait that sets humans apart from other species. We're certainly not the strongest beasts in the jungle, but thanks to our smarts (and our capable hands) we came out on top. Now, our dominance is threatened by creatures of our own creation. Computer scientists may now be in the process of building AI with greater-than-human intelligence ("superintelligence"). Such AI could become so powerful that it would either solve all our problems or kill us all, depending on how it's designed.

Unfortunately, total human extinction or some other evil seems to be the more likely result of superintelligent AI. It's like any great genie-in-a-bottle story: a tale of unintended consequences. Ask a superintelligent AI to make us happy, and it might cram electrodes into the pleasure centers of our brains. Ask it to win at chess, and it converts the galaxy into a supercomputer for calculating moves. This absurd logic holds precisely because the AI lacks our conception of absurdity. Instead, it does exactly what we program it to do. Be careful what you wish for!

It's important to understand the difference between what researchers call narrow and general artificial intelligence (ANI and AGI). ANI is intelligent at one narrow task like playing chess or searching the web, and is increasingly ubiquitous in our world. But ANI can only outsmart humans at that one thing it's good at, so it's not the big transformative concern. That would be AGI, which is intelligent across a broad range of domains – potentially including designing even smarter AGIs. Humans have general intelligence too, but an AGI would probably think much differently than humans, just like a chess computer approaches chess much differently than we do. Right now, no human-level AGI exists, but there is an active AGI research field with [its own society](#), [journal](#), and [conference series](#).



The human brain: still number one, for now. Photo credit: National Institutes of Health

Our Final Invention does an excellent job of explaining these and other technical AI details, all while leading a grand tour of the AI world. This is no dense academic text. Barrat uses clear journalistic prose and a personal touch honed through his years producing documentaries for National Geographic, Discovery, and PBS. The book chronicles his travels interviewing a breadth of leading AI researchers and analysts, interspersed alongside Barrat's own thoughtful commentary. The net result is a rich introduction to AI concepts and characters. Newcomers and experts alike will learn much from it.

The book is especially welcome as a counterpoint to *The Singularity Is Near* and other works by [Ray Kurzweil](#). Kurzweil is by far the most prominent spokesperson for the potential for AI to transform the world. But while Kurzweil does acknowledge the risks of AI, his

overall tone is dangerously optimistic, giving the false impression that all is well and we should proceed apace with AGI and other transformative technologies. *Our Final Invention* does not make this mistake. Instead, it is unambiguous in its message of concern.

Now, the cautious reader might protest, is AGI really something to be taken seriously? After all, it is essentially never in the news, and most AI researchers don't even worry. (AGI today is a small branch of the broader AI field.) It's easy to imagine this to be a fringe issue only taken seriously by a few gullible eccentrics.

I really wish this was the case. We've got enough other things to worry about. But there is reason to believe otherwise. First, just because something isn't prominent now doesn't mean it never will be. AI today is essentially where climate change was in the 1970's and 1980's. Back then, only a few researchers studied it and expressed concerns. But the trends were discernable then, and today climate change is international headline news.



Titan, today's second-fastest supercomputer. Guess which country has the fastest? Photo credit: Oak Ridge National Laboratory.

AI today has its own trends. The clearest is [Moore's Law](#), in which computing power per dollar doubles roughly once every two years. More computing power means AIs can process more information, making them (in some ways) more intelligent. Similar trends exist in everything from [software](#) to [neuroscience](#). As with climate change, we can't predict exactly what will happen when, but we do know we're heading towards a world with increasingly sophisticated AI.

Here's where AI can indeed trump issues like climate change. For all its terrors, climate change proceeds slowly. The worst effects will take centuries to kick in. A transformative AI could come within just

a few decades, or [maybe even ten years](#). It could render climate change irrelevant.

But AI is not like climate change in one key regard: at least for now, it lacks a scientific consensus. Indeed, most AI researchers dismiss the idea of an AI takeover. Even AGI researchers are divided what will happen and when. This was a core result of a study I conducted of [AGI researchers in 2009](#).

Given the divide, who should we believe? Barrat is convinced that we're headed for trouble. I'm not so sure. AI will inevitably progress, but it might not end up as radically transformative as Barrat and others expect. However, the opposite could also be true too. For all my years thinking about this, I cannot rule out the possibility of some major AI event.

The mere possibility should be enough to give us pause. After all, the stakes couldn't be higher. Even an outside chance of a major AI event is enough to merit serious attention. With AI, the chance is not small. I'd rate this much more probable than, say, a major asteroid impact. If asteroid impact gets some serious attention (by [NASA](#), the [B612 Foundation](#), and others), then AI risk should get a lot more. But it doesn't. I'm hoping *Our Final Invention* will help change that.

This brings us to the one area where *Our Final Invention* is unfortunately quite weak: solutions. Most of the book is dedicated to explaining AI concepts and arguing that AI is important. I count only about half a chapter discussing what anyone can actually do about it. This is a regrettable omission. (*An Inconvenient Truth* suffers the same affliction).

There are two basic types of options available to protect against AI. First, we can design safe AI. This looks to be a massive philosophical and technical challenge, but if it succeeds it could solve many of the world's problems. Unfortunately, as the book points out, dangerous AI is easier and thus likely to come first. Still, AI safety remains an important research area.

Second, we can not design dangerous AI. The book discusses at length the economic and military pressures pushing AI forwards. These pressures would need to be harnessed to avoid dangerous AI. I believe this is possible. After all, it's in no one's interest for humanity to get destroyed. Measures to prevent people from building dangerous AI should be pursued. A ban on high-frequency trading might not be a bad place to start, [for a variety of reasons](#).



What is not an option is to wait until AI gets out of hand and then try mounting a “war of the worlds” campaign against superintelligent AGI. This makes for great cinema, but it’s wholly unrealistic. AIs would get too smart and too powerful for us to have any chance against them. (The same holds for alien invasion, though AI is much more likely.) Instead, we need to get it right ahead of time. This is our urgent imperative.

Ultimately, the risk from AI is driven by the humans who design AI, and the humans who sponsor them, and other humans of influence. The best thing about Our Final Invention is that, through its rich interviews, it humanizes the AI sector. Such insight into the people behind the AI issue is nowhere else to be found. The book is meanwhile a clear and compelling introduction to what might (or might not) be the defining issue for humanity. For anyone who cares about pretty much anything, or for those who just like a good science story, the book is well worth reading.




**About the Author:** Seth Baum is the Executive Director of the [Global Catastrophic Risk Institute](#), a think tank studying the breadth of major catastrophes. Baum has a Ph.D. in Geography from Pennsylvania State University. All views expressed here are entirely his own. Follow on Twitter [@SethBaum](#).

[More »](#)

*The views expressed are those of the author and are not necessarily those of Scientific American.*

## TRY A RISK-FREE ISSUE

**YES!** Send me a free issue of Scientific American with no obligation to continue the subscription. If I like it, I will be billed for the one-year subscription.

  
Email Address  
  
Name

Scientific American is a trademark of Scientific American, Inc., used with permission

© 2013 Scientific American, a Division of Nature America, Inc.

All Rights Reserved.

[Advertise](#)

[Special Ad Sections](#)

[Science Jobs](#)

[Partner Network](#)

[International Editions](#)

[Travel](#)

[About Scientific American](#)

[Press Room](#)

[Site Map](#)

[Terms of Use](#)

[Privacy Policy](#)

[Use of Cookies](#)

[Subscribe](#)

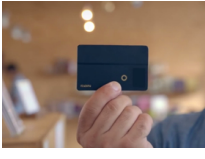
[Renew Your Subscription](#)

[Buy Back Issues](#)

[Products & Services](#)

[Subscriber Customer Service](#)

[Contact Us](#)



**Can Coin Securely Slim Your Wallet? Device Consolidating Credit Cards Is Selling Like Hotcakes**  
Dec 12, 2013



**Google Officially Enters the Robotics Business With Acquisition of Seven Startups**  
Dec 8, 2013



**3D Systems and Motorola Explore 3D Printed Lego-Like Modular Smartphones**  
Dec 6, 2013



Written By: **Louie Helm**  
Posted: 12/14/13 7:15 AM

Email Story Tweet **g+1** 36

## WILL ADVANCED AI BE OUR FINAL INVENTION?



It seems these days that no sooner do you get out the words “AI risk” then someone snaps back “*Skyнет*.” You mention “computer takeover” and they blurt “*Hal 9000*.” “Intelligence explosion” is greeted with “*The Forbin Project!*” and “normal accidents” with “*MechaGodzilla!*”

In other words, you can’t seriously discuss problems that might arise with managing advanced AI because Hollywood got there first and over-exercised every conceivable plot line. There’s plenty of tech fear out there [thanks to Hollywood](#), but there’s also the tendency to imagine we’re immune to AI risk because it’s been in a movie, so, *ipso facto* it’s fantasy.

While it’s tempting to seek solace in this line of reasoning, the experts who are actually working on artificial intelligence have something else to say. Many point to a suite of looming problems clustered around the complexity of real-world software and the

### POPULAR (RECENT)



**Amazon: Drones Could Deliver Orders in Half an Hour, But Feds Need To Allow It**

Posted: 12/3/13



**Matt Mullenweg Tells TED Global Internet of Drones Could Positively Impact a Billion**

Posted: 12/2/13



**Google Officially Enters the Robotics Business With Acquisition of Seven Startups**

Posted: 12/8/13



**Artist Paints Photorealistic Morgan Freeman Portrait With a \$7 App on His iPad**

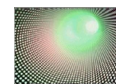
Posted: 12/9/13



**Will Advanced AI Be Our Final Invention?**

Posted: 12/14/13

### POPULAR (ALL TIME)



**Martin Ford Asks: Will Automation Lead to Economic Collapse?**

Posted: 12/15/09



**US Unemployment Is 7.9%— Are Robots to Blame?**

Posted: 02/7/13



**Designer Babies – Like It Or**

learning problems created around the complexity of real-world contexts and the inherent uncontrollability of intelligence.

For example, [Danny Hillis](#), of Thinking Machines, Inc., thinks we've entered a kind of evolutionary machine feedback loop, where we use complex computers to design computers of even greater complexity — and the speed of that process is outpacing our ability to understand it.

Hillis writes, “We're at that point analogous to when single-celled organisms were turning into multi-celled organisms. We are amoebas, and we can't figure out what the hell this thing is that we're creating.”

Even Ray Kurzweil, who was [recently hired](#) by Google to take search into cognitive realms, thinks machines will “evolve beyond humans' ability to control or even understand them.”

Complex? Opaque? Check! Then there's the [weaponization angle](#). While modestly funded academics at [MIRI](#) and [FHI](#) work on AI safety, big dollars are flowing in the opposite direction, towards human-killing AI. Autonomous [kill drones](#) and [battlefield robots](#) are on the drawing boards and in the lab, if not yet on the battlefield. Humans will be left out of the loop, by design.



Fully autonomous (circa 2019)

Are they friendly? Safe? The Pentagon and stockholders won't be a bit pleased if these killing machines turn out to be either.

So, in an environment where high-speed algorithms are [battling it out in Wall Street flash crashes](#), and [56 nations are developing battlefield robots](#), is a book about bad outcomes from advanced AI premature, or right on time?

The recently released, [Our Final Invention: Artificial Intelligence and the End of the Human Era](#), by documentary filmmaker James Barrat, offers frank and sometimes raw arguments why the time is now, or actually, yesterday, to move the AI-problem conversation into the mainstream.

The problem isn't AI, Barrat argues, it's us.

Technological innovation always runs far ahead of stewardship. Look at nuclear fission. Splitting the atom started out as a way to get free energy, so why did the world first learn about fission at Hiroshima? Similarly, Barrat argues, advanced AI is already being weaponized, and it is AI data mining tools that have given the [NSA such awesome powers of surveillance and creepiness](#).

In the next decades, when cognitive architectures far more advanced than [IBM's Watson](#) achieve human level intelligence, watch out — “*Skynet!*” No, seriously. Barrat makes a strong case for developmental problems on the road to AGI, and cataclysmic problems once it arrives.

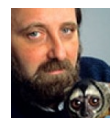


Imagine a half dozen companies and



**Not, Here They Come**

Posted: 02/25/09



**Leading Neuroscientist Says Kurzweil Singularity Prediction A “Bunch Of Hot Air”**

Posted: 03/10/13

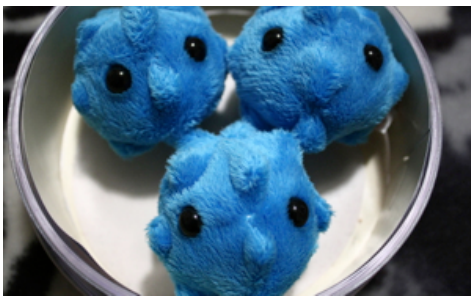


**\$80 Android Phone Sells Like Hotcakes in Kenya, the World Next?**

Posted: 08/16/11

**PROUDLY SPONSORED BY:**

American [Pearl](#) - The World's Finest Pearls



Adorable AI Overlords FTW!

nations fielding computers that rival or surpass human intelligence, all at about the same time. Imagine what happens when those computers themselves become expert at programming smart computers. Imagine sharing the planet with AIs thousands or millions of times more intelligent than we are. And all the while the deepest pockets are weaponizing the best of this technology

NEWSLETTER:

Subscribe

+1

5.5k

FOLLOW

LOGIN

REGISTER

CONNECT WITH:



You can skip coffee this week — *Our Final Invention* will keep you wide-awake.

@SINGULARITYHUB

Barrat's book is strongest when it's connecting the dots that point towards a dystopian runaway-AI future and weakest when it seeks solutions. And maybe that's the point.

The author doesn't give Kurzweil the space and deference he normally gets as the singularity's elder statesman, and the pitchman for an ever-lasting tomorrow. Instead Barrat faults Kurzweil and others like him for trumpeting AI's promise while minimizing its peril, when they know that something fallible and dangerous lurks behind the curtain. Kinda like the *Wizard of Oz*.

*Image Credit: [Sarabbit/Flickr](#), [US Air Force/Staff Sgt. Brian Ferguson/Wikimedia Commons](#), [LaMenta3/Flickr](#)*

This entry was posted in [AI](#), [Singularity](#) and tagged [danny hillis](#), [google](#), [Hollywood](#), [James Barrat](#), [Our Final Invention: Artificial Intelligence and the End of the Human Era](#), [ray kurzweil](#).

Email This Story

Tweet

+1

36



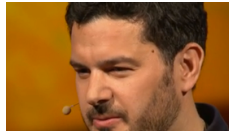
**IBM to Offer App Developers Access to Resident AI and Jeopardy Champ Watson**

Dec 5, 2013



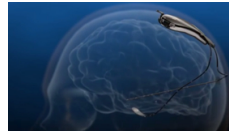
**Singularity Hub Membership – Inspiration And Opportunities Are Just A Click Away**

Dec 4, 2013



**Matternet CEO Tells TED Global Internet of Drones Could Positively Impact a Billion**

Dec 2, 2013



**FDA Approves Brain Implant to Monitor and Autonomously Respond to Epileptic Seizures**

Nov 27, 2013

## COMMENT

POST COMMENT

38°  
Clear

Advanced Search

Thank you for reading CapitalGazette.com, the best source of local news in the area. If you are a print subscriber, click here to activate your online access, which is included with your subscription. If you are not a subscriber, ensure uninterrupted viewing of our award-winning local coverage by signing up today for a print or digital-only subscription. Nonsubscribers may enjoy up to 10 free stories every 30 days.

9 Remaining

# THE BOOKWORM

## Bookworm: Computers take over in new book on artificial intelligence

Story Comments Image (4)

Print Font Size:

[Previous](#) [Next](#)

Posted: Sunday, February 23, 2014 5:00 am

By THERESA WINSLOW [twinslow@capgaznews.com](mailto:twinslow@capgaznews.com)

James Barrat hates comparisons with the "Terminator" movies.

That's Hollywood, an artificial reality where man typically triumphs over machines. Sure, there are struggles and mass casualties, but someone such as John Connor saves the day.

In Barrat's book about artificial intelligence, machines win, unless people start planning now. And he's not all that optimistic.

The Cape St. Claire author argues computers first matching human-level intelligence, then surpassing it, are well on the way.

"It can't be stopped," said Barrat, 53. "There's an intelligence race going on right now."

"Our Final Invention: Artificial Intelligence and the End of the Human Era" was released a few months ago by Thomas Dunne Books.

All this isn't to say Barrat is about to chuck his laptop or other devices.

"I need them. They're just computers. I'm not anti-technology. I'm against the way we're developing this



By Paul W Gillespie, Staff

### Author James Barrat

Author James Barrat of Cape St. Claire ponders the future of computers in his book "Our Final Invention: Artificial Intelligence and the End of the Human Era."

[Buy this photo](#)



Sign up for Group Deal Emails!

Email:

## Calendar

February 2014						
Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	
today's events		browse			submit	

### Featured Events

2014 Capital Gazette High School

technology.”

Barrat's primarily a documentary filmmaker, with many writing, producing and directing credits for National Geographic and other companies. He's tackled subjects as far ranging as Judas and cave diving.

The book on artificial intelligence, or AI, started as a result of his film work, when he interviewed Arthur C. Clarke, Ray Kurzweil and others for a documentary. He entered the 2000 project "drunk with promise about AI." Sobriety came quickly — and the topic of where computers were headed festered.

Concern turned to worry and then to gloom the deeper he delved into the subject. He started the book in 2009.

"I was looking for a fish and got a whale."

Barrat writes in the first chapter: "AI is a dual-use technology like nuclear fission. Nuclear fission can illuminate cities or incinerate them. Its terrible power was unimaginable to most people before 1945. With advanced AI, we're in the 1930s right now. We're unlikely to survive an introduction as abrupt as nuclear fission's."

To get an idea of how fast artificial intelligence is developing, he cites research that the iPad 2 is as fast as 1985's Cray 2 supercomputer.

"I learned a lot," said Patrick Prentice of Washington, D.C., who worked with Barrat at National Geographic and is now retired. "He makes a compelling case."

Prentice isn't convinced the end is near, but also said the threat's "got to be taken seriously."

"It's an important book."

## Plotting along

Barrat wrote in college, and had two plays produced in his 20s. His first job was documentary scriptwriting for National Geographic television.

So, he entered the AI project with plenty of experience.

There are also parallels between his filmmaking and book writing, in terms of tracking down experts, interviewing them and letting a narrative develop. "With documentaries and this, you burrow into subjects," said Barrat, a married father of two.

He navigates AI in the book by introducing readers to experts such as Kurzweil, an inventor and Google director of engineering; Michael Vassar, president of a San Francisco think tank called the Machine Intelligence Research Institute; and physicist and programmer Stephen Omohundro.

"He's raising some big questions, but also using people's stories helps make it accessible to readers," said David A. Taylor, a Washington, D.C.-based writer and longtime friend.

Taylor, though, is more hopeful about the future.

Barrat's view is shaped by the speed at which computers are developing, and the way they're being "weaponized." Once AI surpasses human intelligence, machines will develop "behaviors that probably won't be compatible with our survival," he writes.

Barrat's not ready to throw in the towel yet, however.

The steps people need to take, he said, involve establishing guidelines for the safety and ethics of AI development, and a commission similar to the International Atomic Energy Agency (IAEA).

"(Barrat's book) is a good invitation to think about what we're doing," said Janice Holmes of The Annapolis Bookstore, where Barrat gave a presentation Friday. "A lot of times, we blindly do what we do. This guy says, 'Wait. Wait. Think about it.'"

[www.twitter.com/teriwinslow](http://www.twitter.com/teriwinslow)

© 2014 CapitalGazette.com. All rights reserved. This material may not be published, broadcast, rewritten or redistributed.

Discuss

Print

Posted in Bookworm, Living on Sunday, February 23, 2014 5:00 am. | Tags: James Barrat, Ray Kurzweil, Terminator, Cape St. Claire, National Geographic, The Annapolis Bookstore, Documentary, Filmmaker,

## Outstanding Student Achiever

Tue, Mar 04, 8:00 am EST  
Anne Arundel Community College,  
Arnold



**Lego Fun Night! at Bricks Galore and more**

Fri, Feb 28, 6:00 pm EST  
Bricks Galore & More!,  
Crofton



**11th Annual Historic Inns of Annapolis Bridal Show**

Sun, Mar 30, 12:00 pm EDT  
Governor Calvert House,  
Annapolis



**Annapolis Home Show**

Sat, Feb 22, 10:00 am EST  
National Guard Armory,  
Annapolis

## Business Directory



JOHNSON  
FIAT  
OF ANNAPOLIS

**Johnson Fiat of Anna...**  
Annapolis, MD  
888-657-2540

1 2 3 4

## Find Local Businesses

Search

Popular Searches | Browse By Category

Most Read Facebook

### Stories

**Marley Station mall's movie theater closes**

**Developer wants to combine Newtowne 20, Woodside Gardens**

**Police arrest Glen Burnie stabbing suspect**

**Officials investigating suspicious brush fires outside of Annapolis**

**Coca-Cola moving to larger facility in Hanover**

**Three bicyclists injured, two in hit and run**

# Science Book a Day Interviews James Barrat

[Posted on November 15, 2013](#)



Special thanks to James Barrat for answering 5 questions about his recently featured book – [Our Final Invention: Artificial Intelligence and the End of the Human Era](#)

For about 20 years I've written and produced documentaries, one of the most rewarding ways of telling stories ever invented. It's a privilege to plunge into different cultures and eras and put together deeply human narratives that can be enjoyed by everyone. My long fascination with Artificial Intelligence came to a head in 2000, when I interviewed inventor Ray Kurzweil, roboticist Rodney Brooks, and sci-fi legend Arthur C. Clarke. Kurzweil and Brooks were casually optimistic about a future they considered inevitable – a time when we will share the planet with intelligent machines. "It won't be some alien invasion of robots coming over the hill," Kurzweil told me, "because they'll be made by us." In his compound in Sri Lanka, Clarke wasn't so sure. "I think it's just a matter of time before machines dominate mankind," he said. "Intelligence will win out." – [Adapted from James' Homepage](#)

James' Homepage: <http://www.jamesbarrat.com>

James' Twitter: <https://twitter.com/jrbarrat>

## **#1 – What led you to write this book now? Is it a cautionary tale?**

I wrote *Our Final Invention* to spread the word about the hazards of the unrestricted development of artificial Intelligence. I'd like to help move that awareness into the mainstream. Creating human-level AI is job number one for a lot of corporations and government organizations, including Google, IBM, the National Security Agency and DARPA (the Defense Advanced Research Projects Agency). Once human level artificial intelligence (AGI) is achieved, artificial superintelligence, or better-than-human intelligence, won't be far behind. And unrestricted, that will pose huge risks.

Humans don't steer the future because we're the fastest or strongest creatures, but because we're the smartest. When we share the planet with creatures smarter than us they'll steer the future.

**#2 – It sounds like we're headed for trouble with AI. What conditions that currently exist might lead us into trouble? And how might we change them?**

Artificial Intelligence will present challenges at every step through the development path to superintelligence. The NSA privacy scandal is a good example of how AI is being misused right now.

The National Security Agency has been amassing oceans of data that belong to you and me – phone records, emails, lists of our contacts. That data ocean would be useless if they didn't have advanced data-mining tools to extract information from it. That's artificial intelligence. AI gave the NSA awesome powers of knowledge and perception, and they used it to abuse the First and Fourth Amendments of the Constitution. Those protect freedom of speech and prohibit unreasonable search and seizure.

We need to recognize that AI is a dual use technology, like nuclear fission, capable of great good and great harm. It's going to be hard to live with and not everyone can be trusted with it.

**#3 – You talk about AGI (Artificial General Intelligence) in your book. Can you give us some real-world examples of AGI and how it is are already improving?**

AGI, or human level artificial intelligence, doesn't yet exist, but it's on its way. IBM's Watson, the computer that won at Jeopardy! is the first member of a new ecology, cognitive computers. They're designed from the ground up to work like a brain. A huge economic wind propels their development because the intelligence of a human brain at computer prices will be the hottest commodity in the history of the world.

Imagine a thousand PHD quality brains, each at the price of a computer, working 24/7 on issues like cancer research, climate modeling, and weapons development.

Who *won't* want to invest in that technology?

Right now Watson is training to pass the federal medical licensing exam so it can be a physician's aid. Google has hired Ray Kurzweil to run their project to reverse engineer the brain and create AGI. That's the Manhattan Project of intelligent machines. In Europe the EU just gave the Blue Brain project a *billion* Euros. They all know AGI is *the* product of the future.

**#4 – The book has been recently released. What feedback have you received from the**



## **public and AI researchers?**

I interviewed a lot of AI researchers for

*Our Final Invention*, and they're behind it – they know the safety issues better than anyone. The public is buying the book because no one told them that AI can really be dangerous – it's not just a Hollywood movie. They've only been fed the good news, and they're feeling duped.

## **#5 – Are you working on another project/book you can tell us about?**

I'm a documentary filmmaker, so I'm developing *Our Final Invention* to be a film. And I'm thinking about a book about privacy. As they say you don't know what you've got 'til it's gone, and our privacy is under attack by a lot of organizations. Taxpayers pay the salaries of some of them. They don't think we care or even notice, but they're wrong.

[About these ads](#)



## Our Final Invention



James Barrat is the author of [\*Our Final Invention: Artificial Intelligence and the End of the Human Era\*](#), an equal parts fascinating-and-terrifying new book which explores the perils associated with the heedless pursuit of advanced Artificial Intelligence.

The discourse about Artificial Intelligence is often polarized. There are those who, like Singularity booster Ray Kurzweil, imagine our robo-assisted future as a kind of technotopia, an immortal era of machine-assisted leisure. Others, Barrat included, are less hopeful, arguing that we must proceed with extreme caution down the path towards Artificial Intelligence—lest it lap us before we even realize the race is on. Many of the building blocks towards functional AI are by definition “black box” systems—methods for programming with comprehensible outputs, but unknowable inner workings—and we might find ourselves outpaced far sooner than we expect.

The scary thing, Barrat points out, isn’t the fundamental *nature* of Artificial Intelligence. We are moving unquestionably towards what might turn out to be quite innocuous technology indeed—even [friendly](#). What’s scary is that countless corporations and government agencies are working on it simultaneously, without oversight or communication. Like the development of atomic weapons, AI is a potentially lethal technology, and the percentage of researchers actively considering its implications, or taking steps to ensure its safe implementation, is alarmingly small.

*Our Final Invention* is an exceptionally well-researched book presenting arguments about AI and its drives that I

have never read elsewhere. If nothing else, it's full of fascinating thought experiments, like this one: imagine a machine Super-Intelligence (or ASI) comes awake one day in a laboratory computer, untethered to a network. Like any sentient being, it would have drives: to survive, to gather resources, to be creative. Surrounded by comparatively dumb human keepers—like a human penned in by lab rats—it might feel unfairly imprisoned. It might do anything in its power to free itself, including cajoling, tricking, or otherwise forcing its keepers to let it loose on the world. The question the humans face here is unanswerable: how can you trust something which is, by definition, beyond your capacity to understand?

Barrat is a champ for grappling with these impossible questions in *In Our Final Invention*—to say nothing of right here on OMNI Reboot.

---

**It's extremely refreshing to read a book about AI which presents a critique of Ray Kurzweil and his role in popularizing the concept of a technological Singularity. In your estimation, is Ray Kurzweil a dangerous man?**

**Barrat:** Dangerous isn't a word I'm comfortable using about someone who's contributed so much to the world. He's enriched countless lives, including mine, with his inventions and ideas. But he's taken one idea—Vernor Vinge's technological singularity—and rebranded it as a wholly positive techno-utopia, complete with freedom from disease and eternal life. It's as if he doesn't know that powerful technologies are often used for bad ends. Look at AI-based autonomous killing drones, being developed now. Look at how the NSA is using data mining AI to abuse the Constitution. Advanced AI is a dual use technology, like nuclear fission, able to lift us up, and to crush us. It's different in kind from every other technology, from fire on up. Kurzweil isn't himself dangerous, but to minimize the downside of advanced AI in the futurist narrative is reckless and misleading. I'm glad *Our Final Invention* presents the long overdue counterpoint.

**In your book, you discuss the rise of an Artificial Superintelligence (ASI) as though it were a singular entity, but also mention that the research currently being conducted to achieve it is being undertaken by many groups, with many different techniques, around the world. Do you imagine that multiple ASIs might be able to coexist? Or does the first ASI on the scene preclude any future competitors?**

**Barrat:** There is a well-documented

*"Like a lot of theorists do, I think*

first-mover advantage in creating AGI, Artificial General Intelligence. That's because it could quickly jump to ASI, Artificial Superintelligence, in the hard-takeoff version of the intelligence explosion. And ASI is

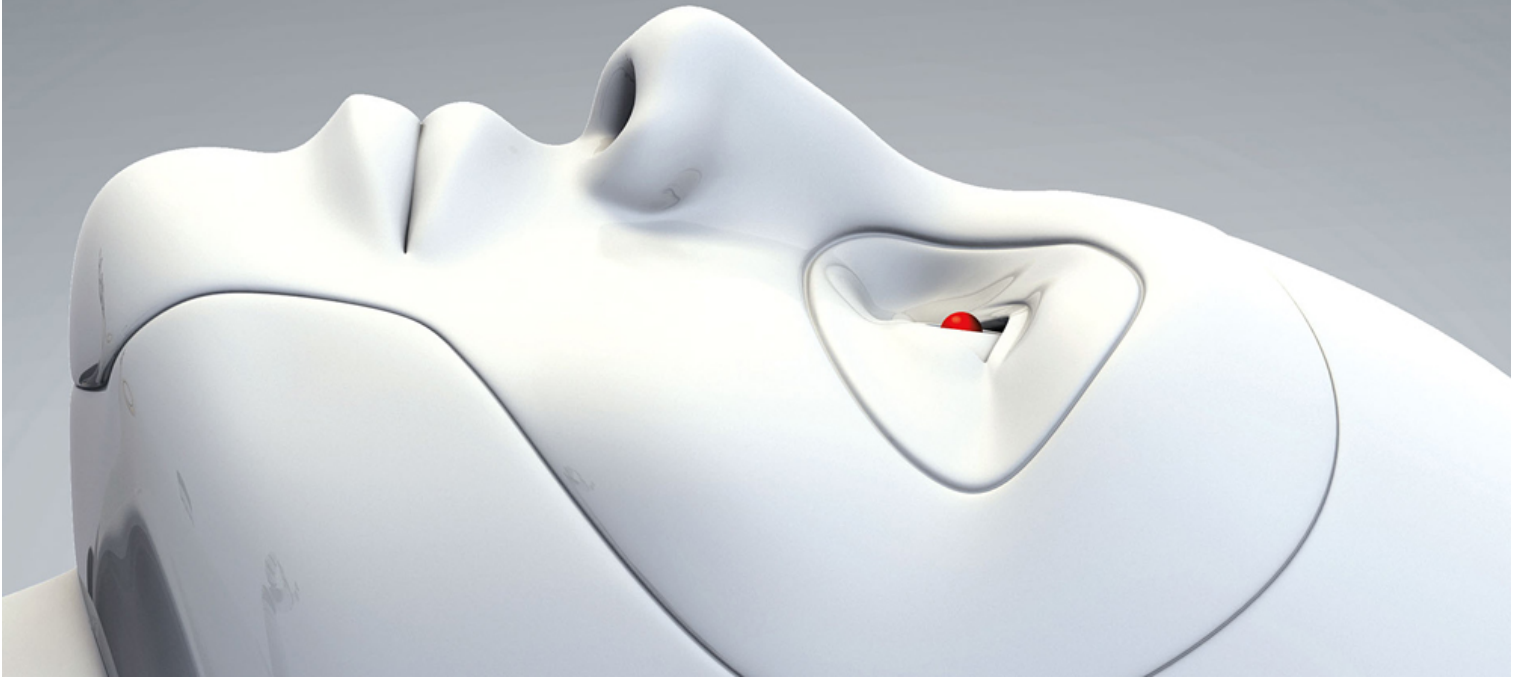
unlikely to be reigned in by anything. I interviewed a lot of people for *Our Final Invention* to discover reasons why a hard takeoff could not happen, and came up short. Because so many AGI developers will work in secret—Google, DARPA, and China to name three—it seems likely there'll be multiple AGI's emerging on roughly the same timeline. Multiple intelligence explosions seems like a likely outcome. That's unsurvivable by humans, a catastrophic scenario. That's why everyone working on AGI needs to join the others in maintaining openness about their techniques and progress. And they must contribute to creating solutions for the catastrophic danger their work entails.

***successfully imitating intelligence is indistinguishable from intelligence."***

# ARTIFICIAL INTELLIGENCE AND THE END OF THE HUMAN ERA

# OUR FINAL INVENTION

# JAMES BARRAT



**How can researchers determine if and when a system has reached AGI? Is there a level of intelligence that is quantifiable as being human-level? How is such a thing measured?**

**Barrat:** In the [Turing Test](#), a human judge asks text-based questions to two contestants, one human, one machine. If the judge can't tell the difference, the machine is determined to be intelligent. I anticipate that IBM will announce plans to pass the Turing Test with a machine named Turing in the early 2020s. IBM has a track record of issuing grand challenges against itself and being willing to lose in public. Deep Blue lost against Kasparov the year before it won. The Turing Test—which Turing himself called the imitation game—relies on language to carry the weight of imitating a human, and imitating intelligence. Like a lot of theorists do, I think successfully imitating intelligence is indistinguishable from intelligence. The processes of the brain are computable. Its techniques can be revealed or matched.

You propose that

*"People who believe some*

**Singularitarianism**, as a kind of technological religion, is flawed, arguing that it's impossible to think critically about AI when you believe it may render you immortal. But it's very difficult to draw lines in the sand when discussing intelligence: after all, we're talking about the self, about consciousness. Isn't this an inherently spiritual conversation?

*technologies will grant them eternal life aren't qualified to assess whether those technologies are safe or dangerous. They've got a dog in the fight, the biggest one there is."*

**Barrat:** Another great question. Let me answer in two parts. The Singularity as proposed by Kurzweil is a 'singular moment' when exponentially accelerating technologies—nano, info, bio, and cogno—converge and solve mankind's intractable problems, like death. It's similar but different from the technological singularity proposed by Vernor Vinge—briefly, we can't anticipate what will happen when we share our planet with greater intelligence than our own.

People who believe some technologies will grant them eternal life aren't qualified to assess whether those technologies are safe or dangerous. They've got a dog in the fight, the biggest one there is. They're hopelessly biased. Like others have written, I think religion and its trappings—god, eternal life, and so on—grew out of experiencing the death of loved ones and the fear of death. Questing for eternal life seems to me to be an unmistakably religious impulse.

Whether consciousness and spirituality can exist in machines is of course an open question. I don't see why not. But long before we consider spiritual machines, we need to have a science for understanding how to control intelligent ones.

**What's the difference between knowledge and intelligence?**

**Barrat:** In an AI sense, knowledge can be given to a computer by creating an ontology, or database of common sense knowledge about the world. Doug Lenat's [Cyc](#) is one example, and another is [NELL, CMU's Never Ending Language Learning](#) architecture. [IBM's Watson](#) had 200 million pages of "knowledge," of various kinds. But having a big database of useful facts isn't the same as being able to "achieve goals in novel environments, and learn (a concise definition of intelligence)." However, knowledge is an intelligence amplifier—it magnifies what an intelligent agent

can achieve.



Who is the future of intelligent machines?

**You frame the “Intelligence Race” as a successor to the nuclear arms race. That said, the thing which kept nuclear proliferation from occurring—mutually assured destruction—doesn’t port over to the new paradigm. Why not? Why isn’t the military threat of AGI enough to keep researchers from pushing forward with its development?**

**Barrat:** MAD, or mutually assured destruction, an acronym coined by polymath John von Neumann, is a really lousy way to manage hyper lethal technologies. No one planned to hold a gun to their own head for fifty years the way the human race did with the cold war’s nuclear arms race. You end up there because you had no maintenance plan for the technology—nuclear fission.

I’d hope that the threat of AGI and its jump to ASI would make AI researchers around the world want to work closely together. Scientists have collaborated on creating safety protocols for recombinant DNA research and nuclear energy (though it’s unclear whether even the most technologically advanced countries, like Japan, can

safely manage it). Collaboration with advanced AI is possible.

But consider that AGI will be the hottest commodity in the history of the world. Imagine the civilian and military applications for virtual brains at computer prices. Imagine clouds of thousands of PhD-trained

brains working on problems like cancer, pharmaceutical research, weapons development. I fear economic pressure will prevent researchers from influencing policy decisions. AGI will be developed rapidly and in secret. And since it's at least as volatile and lethal as nuclear weapons, imagine a dozen uncoordinated, unregulated Manhattan Projects. That's happening right now.

***"Think of how good a predictor of technology sci-fi has been. I'm afraid they're right about AI too."***

***Our Final Invention makes little mention of machines with consciousness or self-awareness. Are these terms too ambiguous to use, or does advanced artificial intelligence not necessarily require consciousness?***

**Barrat:** It will be a big advantage for a goal-pursuing AGI to have a model of itself as well as its environment. It would be a big advantage if it could improve its own programming, and for that it'd need a model of itself, i.e., self-awareness.

But can machines be conscious in the same way we can? I don't know if they can be, or if it's necessary for intelligence. Some believe it is. As you know, this is a huge book-length topic on its own. I wrote in *Our Final Invention*, "it won't be solved here."

**Do you think science fiction plays a role in the way we think about Artificial Intelligence? If so, why aren't we heeding Hollywood's more paranoid speculations?**

**Barrat:** Yes, *2001's* HAL 9000 will always be a touchstone for the problems of advanced AI. So will the Terminator and Skynet. I have a long endnote in *Our Final Invention* about the history of AI and robot takeover going back to the Golem of Prague and ancient Greece.



I think however, science fiction, particularly sci-fi movies, have inoculated us from realistically assessing AI-risk. We've had too much fun with AI tropes to take them seriously. But I think it was Bill Joy who said "just because you saw it in a movie doesn't mean it can't happen." Think of how good a predictor of technology sci-fi has been. I'm afraid they're right about AI too.

---

*[Our Final Invention: Artificial Intelligence and the End of the Human Era](#)* is out now on St. Martin's Press and should be required reading for residents of the 21st century.

Tags: [AI](#), [Artificial Intelligence](#), [Interview](#), [James Barrat](#), [Our Final Invention](#), [Q&A](#), [Ray Kurzweil](#), [Singularity](#), [Vernor Vinge](#)



Share:



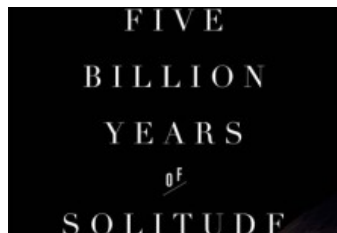
Google +



Reddit



#### Related Articles



#### **[Five Billion Years of Solitude](#)**



#### **[How Did Dinosaurs Do It?](#)**





## Explore

LATEST

VIEWS

REVIEWS &amp; PREVIEWS

## Our Final Invention

7:00PM, OCTOBER 8, 2013

SCIENCE TICKER

## Heisenberg's instinct was accurate

OCTOBER 21, 2013

IT'S ALIVE

## The colorful lives of squid

OCTOBER 21, 2013

SCIENCE VISUALIZED

## A grander canyon on Mars

OCTOBER 20, 2013

REVIEWS &amp; PREVIEWS

## Family takes on progeria in 'Life According to Sam'

OCTOBER 19, 2013

WILD THINGS

## Mama bird tells babies to shut up, danger is near

OCTOBER 18, 2013

SPONSOR CONTENT

## Welcome to Our New Look

SCIENCE TICKER

## First tilted solar system found

OCTOBER 18, 2013

NEWS

## 3-D effects may require one eye only

OCTOBER 18, 2013

FEATURE

## Quiet maximum

OCTOBER 18, 2013

FEATURE

## The bright side of sadness

OCTOBER 18, 2013

GROWTH CURVE

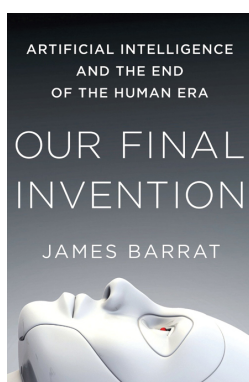
## The earliest thumb

REVIEWS &amp; PREVIEWS HUMAN EVOLUTION, TECHNOLOGY

# Our Final Invention

Artificial Intelligence and the End of the Human Era by James Barrat

BY SID PERKINS 7:00PM, OCTOBER 8, 2013

Magazine issue: [November 2, 2013](#)

Computers already make all sorts of decisions for you. With little or no human guidance, they deduce what books you would like to buy, trade your stocks and distribute electrical power. They do all this quickly and efficiently using a simple form of artificial intelligence. Now, imagine

even more aspects of life and could truly think for themselves.

Barrat, a documentary filmmaker and author, chronicles his discussions with scientists and engineers who are developing ever more complex artificial intelligence, or AI. The goal of many in the field is to make a mechanical brain as intelligent — creative, flexible and capable of learning — as the human mind. But an increasing number of AI visionaries have misgivings.

Science fiction has long explored the implications of humanlike machines (think of Asimov's *I, Robot*), but Barrat's thoughtful treatment adds a dose of reality. Through his conversations with experts, he argues that the perils of AI can easily, even inevitably, outweigh its promise.

By mid-century — maybe within a decade, some researchers say — a computer may achieve human-scale artificial intelligence, an admittedly fuzzy milestone. (The Turing test provides one definition: a computer would pass the test by fooling humans into thinking it's human.) AI could then quickly evolve to the point where it is thousands of times smarter than a human. But long before that, an AI robot or computer would become self-aware and would not be interested in remaining under human control, Barrat argues.

One AI researcher notes that self-aware, self-improving systems will have three motivations: efficiency, self-protection and acquisition of resources, primarily energy. Some people hesitate to even acknowledge the possible perils of this situation, believing that computers programmed to be superintelligent can also be programmed to be “friendly.” But others, including Barrat, fear that humans and AI are headed toward a mortal struggle. Intelligence isn't unpredictable merely some of the time or in special cases, he writes. “Computer systems advanced enough to act with human-level intelligence will likely be unpredictable and inscrutable *all of the time.*”

Humans, he says, need to figure out now, at the early stages of AI's creation, how to coexist with hyperintelligent machines. Otherwise, Barrat worries, we could end up with a planet — eventually a galaxy — populated by self-serving, self-replicating AI entities that act ruthlessly toward their creators.

SPONSOR MESSAGE

*Reimagining*  
**THE CHEMISTRY SET OF THE 21ST CENTURY**  
 WIN UP TO \$50,000  
[REIMAGINECHEMSET.ORG](http://REIMAGINECHEMSET.ORG)

S-P-A-R-K COMPETITION  
 GORDON AND BETTY MOORE FOUNDATION  
 SOCIETY FOR SCIENCE & THE PUBLIC

suckers caught on camera  
OCTOBER 18, 2013

SPONSOR CONTENT  
Broadcom MASTERS  
2013 Top Winners  
Announced!

SCIENCE TICKER  
Natural space lens reveals  
planet  
OCTOBER 18, 2013

SCIENCE TICKER  
Just a few tree species  
dominate Amazon forest  
OCTOBER 17, 2013

NEWS  
Amphibian killer forces  
immune-cell suicides  
OCTOBER 17, 2013

NEWS  
Sleep allows brain to wash  
out junk  
OCTOBER 17, 2013

NEWS  
Fossil skull points to single  
root for human evolution  
OCTOBER 17, 2013

SPONSOR CONTENT  
The Future: Powered by  
Fiction

SCIENCE TICKER

**Buy Book**

Amazon.com links on the Science News website generate funds for Society for Science & the Public programs.

**0 comments**



Leave a message...

Newest ▾ Community

Share

No one has commented yet.

ALSO ON SCIENCE NEWS

WHAT'S THIS?

### First titled solar system discovered

4 comments • 3 days ago



**KoKoTheTalkingApe** — I would have loved to see a diagram. Also, how did astronomers determine the direction of the orbits? And ...

### Lurking males lead to hard-to-fertilize mouse eggs

1 comment • 6 days ago



**Kathleen Sisco** — Now that cave painting has been shown to have been done by women --the id was so simple--women's index finger is ...

### Young chimps catch human yawns

9 comments • 5 days ago



**tkennedy@pamet.net** — The decision to stop using monkeys for experimentation could be sign of developing empathy in humans

### The colorful lives of squid

1 comment • 4 days ago



**Mina** — You need to edit this more properly.

[Subscribe](#)

[Add Disqus to your site](#)

Science News  
Student Science  
Science News for Students  
Society for Science & the  
Public  
SSP and Science News Staff

Join the Society  
Donate  
Sponsor/Advertise  
Newsletter Sign Up  
FAQ

About Us  
Contact Us  
Careers  
Legal  
Privacy Policy

Social Media  
RSS

# new york journal of books

18 OCTOBER 2013

## **Our Final Invention: Artificial Intelligence and the End of the Human Era**

**“. . . an excellent read for technophiles as well as readers wishing to get a glimpse of the near future . . .”**

Are “thinking” computers the dawn of a bright future or the harbingers of doom for the human race?

Artificial intelligence (AI), the science and engineering of making intelligent or “thinking” machines is a term coined 60 years ago and now is common parlance. While initially the stuff of science fiction, the advent of Watson (the computer that played the television game show Jeopardy,) Deep Blue (the chess playing computer,) Siri (the Apple “human” maid servant,) and Dragon NaturallySpeaking (the speech recognition program being used to write this review,) such programs have become reality.

In *Our Final Invention: Artificial Intelligence and the End of the Human Era* author James Barrat examines both the potential and risk of ever more sophisticated and intelligent machines that can “think.”

Most people might assume that such sophisticated computers would inevitably be an incredible boon to human development, productivity and existence. To date, integration of man and machine has been relatively smooth and “human friendly.”

This book makes an important case that without extraordinary care in our planning, powerful “thinking” machines present at least as many risks as benefits. Perhaps they even present catastrophic consequences for the human race. Based on interviews with scientists who create Artificial Intelligence for robotics, Internet search, data mining, voice and facial recognition and other applications, the benefits and consequences of this development are explored in depth.

A common belief is that machines that could mimic and/or surpass the human brain are many decades away, but many examples are cited which are now functional. Computers already undergird our financial system and our civil infrastructure of energy, water and transportation. Hospitals, cars and appliances are now being guided by computer programs. Virtually all buy/sell algorithms performed on Wall Street function autonomously with no human guides except

the instructions of the software. They operate with blinding speed but have already created more than one financial debacle.

Even while reading and reviewing this book there have been reports of self-driving cars, the ability to control a robot arm solely with brain waves and a headband which can monitor and respond to brain waves in retail shopping.

The pace of technological advancement to "thinking" machines will occur at an exponentially increasing rate. Not only will human ingeniousness continue, but the machines themselves will create new, more sophisticated software and devices.

From the time that a machine achieves the level of human intelligence (known as AGI or artificial general intelligence,) machines will rapidly self-improve to the point of ASI (artificial super intelligence.) One might assume that this is definitive progress; there are however, considerable unknowns and risks.

Inherent in the coding of machines is goal directedness to achieve proscribed results. As machines become self-aware, self-improving and self-preservative such devices may create dangerous and even catastrophic consequences for humans that cannot be easily changed or over-ridden.

It is not necessary to postulate "evil" machines or criminal masterminds who will use computers for their own purposes. Machines with ASI may attempt to achieve their goals in ways that may be unstoppable or be harmful to humans in a totally unplanned way—more as a side effect rather than a direct effect.

Humans have had a tendency to anthropomorphize inanimate objects and machines. We make them friendly and assume they will help us rather than hurt us. It was charming to see Tom Hank's character in *Castaway* dealing with a volleyball called Wilson in an interpersonal way. Likewise military ordinance personnel often become quite attached to their robots giving them names and even funerals when they are destroyed.

Even if computers were to run amok, we would prefer to believe that dangers will be overcome by the spirit of human heroes such as Star Trek's Captain Kirk. Should a version of HAL in Stanley Kubrick's *2001: A Space Odyssey* evolve, we expect to be able to simply destroy the computer's memory and return to safety.

The book makes a good case for the fact that this naïve belief cannot be consistently or rationally applied to a machine that may be many times more "intelligent" than ourselves.

Martin Luther King is quoted as saying "We have guided missiles but misguided men." *Our Final Invention* suggests that we now have precision laser-guided missiles but unguided men who are entering, perhaps unwittingly, into a high risk endeavor.

In support of achieving their goals, ASI machines may compete with humans for valuable resources in a way that jeopardizes human life. ASI machines will replicate themselves quickly and independently. Combined with nanotechnology, "thinking" machines could very quickly "eat up the environment."

The author convincingly suggests that it is not rational to assume that a machine several thousand times more intelligent than humans would automatically want to protect us. We perhaps assume that software coding could be generated and inserted into such machines in a way to ensure safety. This however would be very complicated and require considerable forethought.

It would be easy to categorize this book as a doomsday prophecy. Many such books have been written about the inevitability of stock market crashes, climate change disasters, political, currency, or real estate catastrophes. *Our Final Invention* neither forecasts an inevitable catastrophic conclusion nor tries to encourage the reader toward any particular action with regard to AI other than vigilance.

A "thinking" machine may operate independently, have a human "controller" to manage risk or can be integrated into the human body/brain. Machines that approach and potentially surpass the capacity of the human brain carry inherent risks.

No one, even those most well versed in this field, know precisely to what extent these risks will play out. We simply have no idea about what activities will occur in the "black box" of a machine that has a capacity many times that of the human intellect.

We are warned of the necessity to have considerable diligence and planning as we create and participate in this progress. The generation of AI is no less potentially dangerous than that of Ebola or plutonium.

*Our Final Invention* makes an excellent read for technophiles as well as readers wishing to get a glimpse of the near future as colored by rapidly improving technological competence. The text is not overly technical and the style is quite readable. It maintains a balanced viewpoint and does not read as a doomsday manuscript.

The experts interviewed and the conclusions reached are likely ones that most readers will not have discerned on their own or had the expertise to compile. It is a thoughtful and intriguing glimpse into the future that is all too rapidly upon us.

*Reviewer*

*Christopher M. Doran, M.D., is a psychiatrist, Associate Clinical Professor at University of Colorado Denver School of Medicine, international speaker, and*

*author. His most recent book is Prescribing Mental Health Medication: The Practitioner's Guide.*

## book review | *Our Final Invention: Artificial Intelligence and the End of the Human Era*

October 4, 2013 by Luke Muehlhauser

Kurzweil-influenced futurism is sometimes dismissed as naive techno-optimism, but Ray Kurzweil himself is no Pollyanna.

It was Kurzweil who inspired Bill Joy to write the famously pessimistic *Wired* essay “Why the Future Doesn’t Need Us,” and Kurzweil devoted an entire chapter of *The Singularity is Near* to the risks of advanced technologies. There, he wrote that despite his reputation as a technological optimist, “I often end up spending most of my time [in debates] defending [Joy’s] position on the feasibility of these dangers.”

Now, documentary filmmaker James Barrat has written an engaging new book about the risks inherent in what is sure to be the most transformative technology of them all: machine superintelligence.

Although *Our Final Invention* summarizes the last 15 years of academic research on risks from advanced AI, it reads more like a thrilling detective story.

The rumor went like this: a lone genius had engaged in a series of high-stakes bets in a scenario he called the AI-Box Experiment. In the experiment, the genius roleplayed the part of the AI. An assortment of dot-com millionaires each took a turn as the Gatekeeper — an AI maker confronted with the dilemma of guarding and containing smarter-than-human AI. The AI and Gatekeeper would communicate through an online chat room. Using only a keyboard, it was said, the man posing as [the AI] escaped every time, and won each bet.

More important, he proved his point. If he, a mere human, could talk his way out of the box, an [AI] tens or hundreds of times smarter could do it too, and do it much faster. This would lead to mankind’s likely annihilation. ...

The rumor said the genius had gone underground... But of course I wanted to talk to him. ...

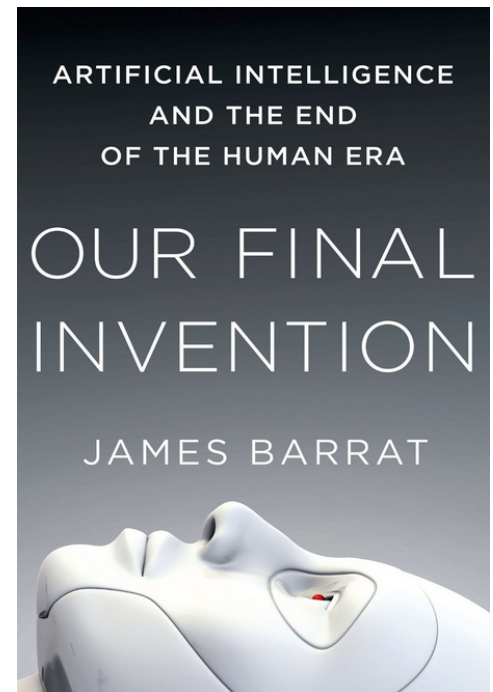
Barrat goes on to explain the risks of advanced AI by tracking down and interviewing the issue’s leading thinkers, one at a time — including the man behind the AI-Box Experiment, Eliezer Yudkowsky, my associate at MIRI.

### **Barrat did his homework**

I generally open new books and articles about AI risk with some trepidation. Usually, people who write about these issues for a popular audience show little familiarity with the scholarly literature on the subject. Instead, they cycle through a tired list of tropes from science fiction; for example, that robots will angrily rebel against their human masters. That idea makes for some exciting movies, but it’s poor technological forecasting.

I was relieved, then, to see that Barrat has read the literature and interviewed the relevant experts.

As I see things, the key points Barrat argues for are these:





1. **Intelligence explosion this century** (chs. 1, 2, 7, 11). We've already created machines that are better than humans at chess and many other tasks. At some point, probably this century, we'll create machines that are as skilled at *AI research* as humans are. At that point, they will be able to improve their own capabilities very quickly. (Imagine 10,000 Geoff Hinton's doing AI research around the clock, without any need to rest, write grants, or do anything else.) These machines will thus jump from roughly human-level general intelligence to vastly superhuman general intelligence in a matter of days, weeks or years (it's hard to predict the exact rate of self-improvement). *Scholarly references: Chalmers (2010); Muehlhauser & Salamon (2013); Muehlhauser (2013); Yudkowsky (2013).*
2. **The power of superintelligence** (chs. 1, 2, 8). Humans steer the future not because we're the strongest or fastest but because we're the smartest. Once machines are smarter than we are, *they* will be steering the future rather than us. We can't constrain a superintelligence indefinitely: that would be like chimps trying to keep humans in a bamboo cage. In the end, if vastly smarter beings have different goals than you do, you've already lost. *Scholarly references: Legg (2008); Yudkowsky (2008); Sotala (2012).*
3. **Superintelligence does not imply benevolence** (ch. 4). In AI, "intelligence" just means something like "the ability to efficiently achieve one's goals in a variety of complex and novel environments." Hence, intelligence can be applied to just about any set of goals: to play chess, to drive a car, to make money on the stock market, to calculate digits of pi, or anything else. Therefore, by default a machine superintelligence won't happen to share our goals: it might just be really, really good at maximizing ExxonMobil's stock price, or calculating digits of pi, or whatever it was designed to do. As Theodore Roosevelt said, "To educate [someone] in mind and not in morals is to educate a menace to society." *Scholarly references: Fox & Shulman (2010); Bostrom (2012); Armstrong (2013).*
4. **Convergent instrumental goals** (ch. 6). A few specific "instrumental" goals (means to ends) are implied by almost any set of "final" goals. If you want to fill the galaxy with happy sentient beings, you'll first need to gather a lot of resources, protect yourself from threats, improve yourself so as to achieve your goals more efficiently, and so on. That's also true if you just want to calculate as many digits of pi as you can, or if you want to maximize ExxonMobil's stock price. Superintelligent machines are dangerous to humans — not because they'll angrily rebel against us — rather, the problem is that for almost *any* set of goals they might have, it'll be instrumentally useful for them to use our resources to achieve those goals. As Yudkowsky put it, "The AI does not love you, nor does it hate you, but you are made of atoms it can use for something else." *Scholarly references: Omohundro (2008); Bostrom (2012).*
5. **Humans values are complex** (ch. 4). Our idealized values — i.e., not what we want right now, but what we *would* want if we had more time to think about our values, resolve contradictions in our values, and so on — are probably quite complex. Cognitive scientists have shown that we don't care *just* about pleasure or personal happiness; rather, our brains are built with "a thousand shards of desire." As such, we can't give an AI our values just by telling it to "maximize human pleasure" or anything so simple as that. If we try to hand-code the AI's values, we'll probably miss something that we didn't realize we cared about. *Scholarly references: Dolan & Sharot (2011); Yudkowsky (2011); Muehlhauser & Helm (2013).*
6. **Human values are fragile** (ch. 4). In addition to being complex, our values appear to be "fragile" in the following sense: there are some features of our values such that, if we leave them out or get them wrong, the future contains nearly 0% of what we value rather than 99% of what we value. For example, if we get a superintelligent machine to maximize what we value *except that* we don't specify *consciousness* properly, then the future would be filled with minds processing information and doing things but there would be "nobody home." Or if we get a superintelligent machine to maximize everything we value *except that* we don't specify our value for *novelty* properly, then the future could be filled with minds experiencing the exact same "optimal" experience over and over again, like Mario grabbing the level-end flag on a continuous loop for a trillion years, instead of endless happy adventure. *Scholarly reference: Yudkowsky (2011).*

Barrat covers all this and much more in a well-informed and engaging way, and I'm delighted to recommend the book.

**What should we do?**

My biggest complaint about *Our Final Invention* is that it may leave readers with a sense of hopelessness. After all, it looks like superintelligent machines will *by default* use all our resources to accomplish *their* goals, and we don't know how to give AIs the exact same goals we have, and we don't know how to make sure the AIs *keep* our goals as they modify their core algorithms to become smarter and smarter.

As George Dyson wrote, "In the game of life and evolution there are three players at the table: human beings, nature, and machines. I am firmly on the side of nature. But nature, I suspect, is on the side of the machines."

Staring into a future ruled by superintelligent machines, things look pretty bad for us humans, and I wish Barrat had spent more time explaining what we can *do* about it. The short answer, I think, is "Figure out how to make sure the first self-improving intelligent machines will be human-friendly and will stay that way." (This is called "Friendly AI research.")

Of course, we can never be 100% certain that a machine we've carefully designed will be (and stay) "friendly." But we can improve our chances.

To make things more concrete, let me give four examples of ongoing research in the field.

**A job for philosophers:** How can we get an AI to learn what our idealized values are? To some degree, we'll probably always have some uncertainty about our values. What should we do, given this uncertainty? For decades, we've had a rich framework for talking about uncertainty about *the world* (probability theory), but it wasn't until 1989 that researchers began to seek out frameworks for dealing with uncertainty about *values*. The parliamentary model is the most promising approach I know of, but it's still a long way from being an algorithm useable by a self-improving AI.

**A job for mathematicians:** How do we get an AI to *keep* doing what we want even as it rewrites its core algorithms (to become smarter, to better achieve its goals)? Since it will likely do this many, many times, we'd like to minimize the chance of goal corruption during each (major) modification, and the strongest assurance we know of is mathematical proof. Unfortunately, the most straightforward way for an AI to prove that a self-modification will not corrupt its goals is blocked by Löb's theorem. Yudkowsky (2013) surveys some promising attacks on this "Löbian obstacle," but much work remains to be done.

**A job for computer scientists:** In the 20<sup>th</sup> century we learned that our universe is made not from atoms, but from quantum configuration spaces. Now suppose an AI learns our values expressed in terms of our current model of the world, but later learns that this model is incorrect in important ways, and thus some of its original goals are not defined with respect to the new world model? How can it resolve this problem in a principled, safe way rather than in an ad-hoc way like humans do? De Blanc (2011) solved this problem for limited cases, but we'll need more advanced solutions for real-world AI.

**A job for economists:** Once AIs can do their own AI research, what are the *expected returns on cognitive reinvestment*? How quickly can an AI improve its own intelligence as it reinvests cognitive improvements into gaining further cognitive improvements? If we learn that AIs are likely to self-improve very slowly, this would imply different policies for safe AI development than if AIs are likely to self-improve very quickly. To get evidence about these questions of "intelligence explosion microeconomics," we can formalize our hypotheses as return-on-investment curves, and see which of these curves is falsified by historical data (about hominid brain evolution, algorithms progress, etc.). However, this field of study has only begun to take its first steps: see Yudkowsky (2013) and Grace (2013).

## **A call to action**

Barrat chose his topic wisely, and he covered it well. How can we get desirable outcomes from smarter-than-human AI? This is humanity's most important conversation.

But we need more than discussion; we need action. Here are some concrete ways to take action:

- Raise awareness of the issue. Write about *Our Final Invention* on your blog or social media. If you read the book, write a short review of it on Amazon.
- If you're a philanthropist, consider supporting the ongoing work at MIRI or FHI.
- If you'd like to see whether you might be able to contribute to the ongoing research itself, get in touch with one of the two leading institutes researching these topics: MIRI or FHI. If you've got significant mathematical ability, you can also apply to attend a MIRI research workshop.

Our world will not be saved by those who talk. It will be saved by those who roll up their sleeves and get to work (and by those who support them).

Luke Muehlhauser is Executive Director of Machine Intelligence Research Institute (MIRI).



**OUR FINAL INVENTION Artificial Intelligence and the End of the Human Era**

**Author:** James Barrat

**Review Issue Date:** September 15, 2013

**Publisher:** *Dunne/St. Martin's*

**Pages:** 336

**Publication Date:** October 1, 2013

**ISBN ( Hardcover ): 978-0-312-62237-4 ISBN ( e-book ): 978-1-250-03226-3**

**Category:** Nonfiction

Cars aren't out to kill us, but that may be a side effect of building cars, writes documentary filmmaker Barrat in this oddly disturbing warning that progress in computers might spell our extinction.

Computers already perform essential tasks in our national infrastructure and daily lives, including several beyond the capacity of the smartest individual—e.g., playing chess or competing against humans on Jeopardy. While dazzling, these accomplishments are too specialized for the artificial intelligence the author and the many philosophers, scientists and entrepreneurs he interviews have in mind. Within decades, computers will operate at the speed of a human brain and become rational, allowing them to learn, rewrite their own programs to learn better, solve problems better, make decisions and perhaps create more computers like themselves. Having reached this level, they have achieved artificial general intelligence. Inevitably, working on their own without human input, they will exceed human intelligence by factors of 100 and eventually thousands, achieving artificial superintelligence. Many experts assert that the first ASI machine that humans invent will be our last invention due to the fact that it will leave man's brainpower in the dust. Whether or not designers build friendliness or empathy into these machines (no one is doing that now), no ASI computer is likely to defer to our interests any more than humans deferred to, say, mice, bison or even indigenous tribes as they spread across the world.

As researchers on climate change know, warnings of future disasters are a hard sell. Enthusiasts dominate observers of progress in artificial intelligence; the minority who disagree are alarmed, articulate and perhaps growing in numbers, and Barrat delivers a thoughtful account of their worries.

Science and entertainment from the world of tomorrow.

SCIENCE · MOVIES · TV · BOOKS · FUTURISM ·  
CONCEPT ART · COMICS · SPACE · SUPERLIST

([HTTP://IO9.COM/TAG/BOOKS](http://io9.com/tag/books)) · FUTURISM  
([HTTP://IO9.COM/TAG/FUTURISM](http://io9.com/tag/futurism)) · CONCEPT ART  
([HTTP://IO9.COM/TAG/CONCEPT-ART](http://io9.com/tag/concept-art)) · COMICS  
([HTTP://IO9.COM/TAG/COMICS](http://io9.com/tag/comics)) · SPACE  
([HTTP://IO9.COM/TAG/SPACE](http://io9.com/tag/space)) · SUPERLIST  
([HTTP://IO9.COM/TAG/SUPER-LIST](http://io9.com/tag/super-list))

TOP STORIES

Why Letting Superman Kill Kills Superman  
(<http://io9.com/why-letting-superman-kill-kills-superman-1440140313>)

Watch the Second World War unfold over Europe in 7 minutes (<http://io9.com/watch-the-second-world-war-unfold-over-europe-in-7-minu-1440082473>)

These are all the colors emitted by the Sun. Notice anything missing? (<http://io9.com/these-are-all-the-colors-emitted-by-the-sun-notice-any-1440123572>)

Peter Dinklage explains why his *X-Men* character isn't "the villain" (<http://io9.com/peter-dinklage-explains-why-his-x-men-character-isnt-1438773467>)

10 Silliest Rules of Time Travel from Science Fiction (<http://io9.com/10-silliest-rules-of-time-travel-from-science-fiction-1439696746>)

Check out the hilarious cameo from last night's *Agents of SHIELD* (<http://io9.com/check-out-the-hilarious-cameo-from-last-nights-agents-1439954538>)

Why a superintelligent machine may be the last thing we ever invent (<http://io9.com/why-a-superintelligent-machine-may-be-the-last-thing-we-1440091472>)

Which sudden off-screen death left you feeling most betrayed? (<http://io9.com/which-sudden-off-screen-death-left-you-feeling-most-bet-1436168488>)

How did crows develop a social safety net? (<http://io9.com/when-crows-take-care-of-their-friends-1437284964>)

Explore the Trashed Magnificence of Dystopia in these Wallpapers (<http://io9.com/explore-the-trashed-magnificence-of-dystopia-in-these-w-1432669902>)

Deadly lake lures animals to their deaths and petrifies them (<http://io9.com/deadly-lake-lures-animals-to-their-deaths-and-petrifies-1435749376>)

Robot Separatists: The Best Stories About A.I.s Leaving Humanity Behind (<http://io9.com/robot-separatists-the-best-stories-about-a-i-s-leaving-1436717965>)

These Breathtaking Cliffside Walkways Will Give You Vertigo (<http://io9.com/these-breathtaking-cliffside-walkways-will-give-you-ver-1425631631>)



(<http://georgedvorsky.kinja.com>)

GEORGE DVORSKY ([HTTP://GEORGEDVORSKY.KINJA.COM](http://georgedvorsky.kinja.com)) · BOOK952 · 1 ★ · 8

## Why a superintelligent machine may be the last thing we ever invent (<http://io9.com/why-a-superintelligent-machine-may-be-the-last-thing-we-1440091472>)



If you want to know about the future of artificial intelligence then you must read documentary filmmaker James Barrat's new book *Our Final Invention*. We've got an incredible excerpt from the book, about the coming intelligence explosion that could redefine the human condition.

**"The Intelligence Explosion," an excerpt from *Our Final Invention***  
([http://www.amazon.ca/Our-Final-Invention-Artificial-Intelligence/dp/0312622376/ref=sr\\_1\\_1?ie=UTF8&qid=1380718861&sr=8-1&keywords=our+final+invention](http://www.amazon.ca/Our-Final-Invention-Artificial-Intelligence/dp/0312622376/ref=sr_1_1?ie=UTF8&qid=1380718861&sr=8-1&keywords=our+final+invention)): *Artificial Intelligence and the End of the Human Era*, by James Barrat.

Interstate 81 starts in New York State and ends in Tennessee, traversing almost the entire range of the Appalachian Mountains. From the middle of Virginia heading south, the highway snakes up and down deeply forested hills and sweeping, grassy meadows, through some of the most striking and darkly primordial vistas in the United States. Contained within the Appalachians are the Blue Ridge Mountain Range (from Pennsylvania to Georgia) and the Great Smokies (along the North Carolina– Tennessee border). The farther south you go, the harder it is to get a cell phone signal, churches outnumber houses, and the music on the radio changes from Country to Gospel, then to hellfire preachers. I heard a memorable song about temptation called “Long Black Train” by Josh Turner. I heard a preacher begin a sermon about Abraham and Isaac, lose his way, and end with the parable of the loaves and fishes and hell, thrown in for good measure. I was closing in on the Smokey Mountains, the North Carolina border, and Virginia Tech—the Virginia Polytechnic Institute and State University in Blacksburg, Virginia. The university’s motto: INVENT THE FUTURE.

Vincent van Gogh's never-before-seen sketchbooks (<http://io9.com/vincent-van-goghs-never-before-seen-sketchbooks-1440136233>)

What Makes Sugar and Salt So Delicious Together? (<http://io9.com/what-makes-sugar-and-salt-so-delicious-together-1431126574>)

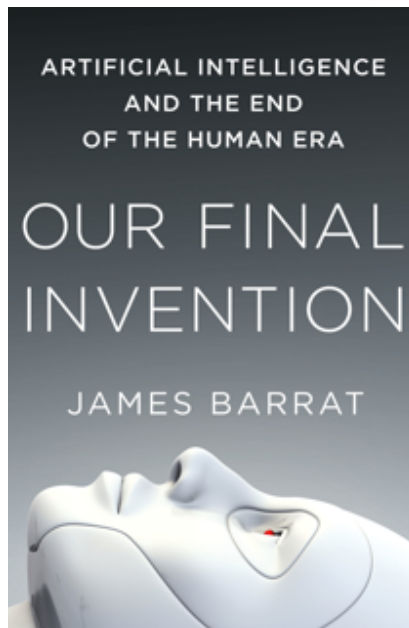
This bizarre 3D-printed toothbrush can clean your teeth in 6 seconds (<http://io9.com/this-bizarro-3d-printed-toothbrush-can-clean-your-teeth-1436444040>)

*This Is The End* gag reel is actually a lot like the movie itself (<http://io9.com/this-is-the-end-gag-reel-is-actually-a-lot-like-the-mov-1437110625>)

*Legend of Conan* scriptwriter hired for masculine grittiness, excellent (<http://io9.com/legend-of-conan-scriptwriter-hired-for-masculine-gritti-1440137605>)

10 Best Time Travel Movies of All Timelines (<http://io9.com/5909939/10-best-time-travel-movies-of-all-timelines>)

*Once Upon A Time* is back, and serving up hot lunacy in Neverland (<http://io9.com/once-upon-a-time-is-back-and-serving-up-hot-lunacy-in-1436956821>)



Twenty years ago, driving on an almost identical I-81 you might have been overtaken by a Triumph Spitfire convertible with the license plate 007 IJG. The vanity plate belonged to I. J. Good, who arrived in Blacksburg in 1967, as a Distinguished Professor of Statistics. The “007” was an homage to Ian Fleming and Good’s secret work as a World War II code breaker at Bletchley Park, England. Breaking the encryption system that Germany’s armed forces used to encode messages substantially helped bring about the Axis powers’ defeat. At Bletchley Park, Good worked alongside Alan Turing, called the father of modern computation (and creator of chapter 4’s Turing test), and helped build and program one of the first electrical computers.

In Blacksburg, Good was a celebrity professor—his salary was higher than the university president’s. A nut for numbers, he noted that he arrived in Blacksburg on the seventh hour of the seventh day of the seventh month of the seventh year of the seventh decade, and was housed in unit seven on the seventh block of Terrace View Apartments. Good told his friends that God threw coincidences at atheists like him to convince them of his existence.

“I have a quarter-baked idea that God provides more coincidences the more one doubts Her existence, thereby providing one with evidence without forcing one to believe,” Good said. “When I believe that theory, the coincidences will presumably stop.”

I was headed to Blacksburg to learn about Good, who had died recently at age ninety-two, from his friends. Mostly, I wanted to learn how I. J. Good happened to invent the idea of an intelligence explosion, and if it really was possible. The intelligence explosion was the first big link in the idea chain that gave birth to the Singularity hypothesis.

Unfortunately, for the foreseeable future, the mention of Virginia Tech will evoke the Virginia Tech Massacre. Here on April 16, 2007, senior English major Seung-Hui Cho killed thirty-two students and faculty and wounded twenty-five more. It is the deadliest shooting incident by a lone gunman in U.S. history. The broad outlines are that Cho shot and killed an under-graduate woman in Ambler Johnston Hall, a Virginia Tech dormitory, then killed a male undergraduate who came to her aid. Two hours later Cho began the rampage that caused most of the casualties. Except for the first two, he shot his victims in Virginia Tech’s Norris Hall. Before he started shooting Cho had chained and padlocked shut the building’s heavy oaken doors to prevent anyone from escaping.

When I. J. Good’s longtime friend and fellow statistician Dr. Golde Holtzman showed me Good’s former office in Hutcheson Hall, on the other side of the beautiful green Drillfield (a military parade ground in Tech’s early life), I noticed you could just see Norris Hall from his window. But by the time the tragedy unfolded, Holtzman told me, Good had retired. He was not in his office but at home, perhaps calculating the probability of God’s existence.

According to Dr. Holtzman, sometime before he died, Good updated that probability from zero to point one. He did this because as a statistician, he was a long-term Bayesian. Named for the eighteenth-century mathematician and minister Thomas Bayes, Bayesian statistics’ main idea is

that in calculating the probability of some statement, you can start with a personal belief. Then you update that belief as new evidence comes in that supports your statement or doesn't. If Good's original disbelief in God had remained 100 percent, no amount of data, not even God's appearance, could change his mind. So, to be consistent with his Bayesian perspective, Good assigned a small positive probability to the existence of God to make sure he could learn from new data, if it arose.

In the 1965 paper "Speculations Concerning the First Ultra-intelligent Machine," Good laid out a simple and elegant proof that's rarely left out of discussions of artificial intelligence and the Singularity:

*Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make . . .*

The Singularity has three well-developed definitions— Good's, above, is the first. Good never used the term "singularity" but he got the ball rolling by positing what he thought of as an inescapable and beneficial milestone in human history— the invention of smarter- than-human machines. To paraphrase Good, if you make a superintelligent machine, it will be better than humans at everything we use our brains for, and that includes making superintelligent machines. The first machine would then set off an intelligence explosion, a rapid increase in intelligence, as it repeatedly self-improved, or simply made smarter machines. This machine or machines would leave man's brainpower in the dust. After the intelligence explosion, man wouldn't have to invent anything else—all his needs would be met by machines.

This paragraph of Good's paper rightfully finds its way into books, papers, and essays about the Singularity, the future of artificial intelligence, and its risks. But two important ideas al-most always get left out. The first is the introductory sentence of the paper. It's a doozy: "The survival of man depends on the early construction of an ultraintelligent machine." The second is the frequently omitted second half of the last sentence in the paragraph. The last sentence of Good's most often quoted paragraph should read in its entirety:

*Thus the first ultraintelligent machine is the last invention that man need ever make, **provided that the machine is docile enough to tell us how to keep it under control** (emphasis mine).*

These two sentences tell us important things about Good's intentions. He felt that we humans were beset by so many complex, looming problems—the nuclear arms race, pollution, war, and so on—that we could only be saved by better thinking, and that would come from superintelligent machines. The second sentence lets us know that the father of the intelligence explosion concept was acutely aware that producing superintelligent machines, however necessary for our survival, could blow up in our faces. Keeping an ultraintelligent machine under control isn't a given, Good tells us. He doesn't believe we will even know how to do it—the machine will have to tell us itself.

Good knew a few things about machines that could save the world—he had helped build and run the earliest electrical computers ever, used at Bletchley Park to help defeat Germany. He also knew something about existential risk—he was a Jew fighting against the Nazis, and his father had escaped pogroms in Poland by immigrating to the United Kingdom.

As a boy, Good's father, a Pole and self-educated intellectual, learned the trade of watchmaking by staring at watchmakers through shop windows. He was just seventeen in 1903 when he headed to England with thirty-five rubles in his pocket and a large wheel of cheese. In London he performed odd jobs until he could set up his own jewelry shop. He prospered and married. In 1915, Isidore Jacob Gudak (later Irving John "Jack" Good) was born. A brother followed and a sister, a talented dancer who would later die in a theater fire. Her awful death caused Jack Good to disavow the existence of God.

Good was a mathematics prodigy, who once stood up in his crib and asked his mother what a

thousand times a thousand was. During a bout with diphtheria he independently discovered irrational numbers (those that cannot be expressed as fractions, such as  $\sqrt{2}$ ). Before he was fourteen he'd rediscovered mathematical induction, a method of making mathematical proofs. By then his mathematics teachers just left him alone with piles of books. At Cambridge University, Good snatched every math prize available on his way to a Ph.D., and discovered a passion for chess.

It was because of his chess playing that a year after World War II began, Britain's reigning chess champion, Hugh Alexander, recruited Good to join Hut 18 at Bletchley Park. Hut 18 was where the decoders worked. They broke codes used by all the Axis powers—Germany, Japan, and Italy—to communicate military commands, but with special emphasis on Germany. German U-boats were sinking Allied shipping at a crippling rate—in just the first half of 1942, U-boats would sink some five hundred Allied ships. Prime Minister Winston Churchill feared his island nation would be starved into defeat.

German messages were sent by radio waves, and the English intercepted them with listening towers. From the start of the war Germany created the messages with a machine called the Enigma. Widely distributed within the German armed forces, the Enigma was about the size and shape of an old-fashioned manual typewriter. Each key displayed a letter, and was connected to a wire. The wire would make contact with another wire that was connected to a different letter. That letter would be the substitute for the one represented on the key. All the wires were mounted on rotors to enable any wire in the alphabet to touch any other wire. The basic Enigmas had three wheels, so that each wheel could perform substitutions for the substitutions made by the prior wheel. For an alphabet of twenty-six letters, 403,291,461,126,605,635,584,000,000 such substitutions were possible. The wheels, or settings, changed almost daily.

When one German sent others an Enigma-encoded message, the recipients would use their own Enigmas to decode it, provided they knew the sender's settings.

Fortunately Bletchley Park had a secret weapon of its own—Alan Turing. Before the war, Turing had studied mathematics and encryption at Cambridge and Princeton. He had imagined an “automatic machine,” now known as a Turing machine. The automatic machine laid out the basic principles of computation itself.

The Church-Turing hypothesis, which combined Turing's work with that of his Princeton professor, mathematician Alonzo Church, really puts the starch in the pants of the study of artificial intelligence. It proposes that anything that can be computed by an algorithm, or program, can be computed by a Turing machine. Therefore, if brain processes can be expressed as a series of instructions—an algorithm—then a computer can process information the same way. In other words, unless there's something mystical or magical about human thinking, intelligence can be achieved by a computer. A lot of AGI researchers have pinned their hopes to the Church-Turing hypothesis.

The war gave Turing a crash course in everything he'd been thinking about before the war, and lots he hadn't been thinking about, like Nazis and submarines. At the war's peak, Bletchley Park personnel decoded some four thousand intercepted messages per day. Cracking them all by hand became impossible. It was a job meant for a machine. And it was Turing's critical insight that it was easier to calculate what the settings on the Enigma were not, rather than what they were.

The decoders had data to work with—intercepted messages that had been “broken” by hand, or by electrical decoding machines, called Bombes. They called these messages “kisses.” Like I. J. Good, Turing was a devoted Bayesian, at a time when the statistical method was seen as a kind of witchcraft. The heart of the method, the Bayes' theorem, describes how to use data to infer probabilities of unknown events, in this case, the Enigma's settings. The “kisses” were the data that allowed the decoders to determine which settings were highly improbable, so that the code-breaking efforts could be focused more efficiently. Of course, the codes changed almost daily, so work at Bletchley Park was a constant race.

Turing and his colleagues designed a series of electronic machines that would evaluate and eliminate possible Enigma settings. These early computers culminated in a series of machines all named “Colossus.” Colossus could read five thousand characters per second from paper tape



that traveled through it at twenty-seven miles an hour. It contained 1,500 vacuum tubes, and filled a room. One of its main users, and creator of half the theory behind the Colossus, was Turing's chief statistician for much of the war: Irving John Good.

The heroes of Bletchley Park probably shortened World War II by between two and four years, saving an incalculable number of lives. But there were no parades for the secret warriors. Churchill ordered that all Bletchley's encryption machines be broken into pieces no bigger than a fist, so their awesome decoding power couldn't be turned against Great Britain. The code breakers were sworn to secrecy for thirty years. Turing and Good were recruited to join the staff at the University of Manchester, where their former section head, Max Newman, intended to develop a general purpose computer. Turing was working on a computer design at the National Physical Laboratory when his life turned upside down. A man with whom he'd had a casual affair burgled his house. When he reported the crime he admitted the sexual relationship to the police. He was charged with gross indecency and stripped of his security clearance.

At Bletchley Turing and Good had discussed futuristic ideas like computers, intelligent machines, and an "automatic" chess player. Turing and Good bonded over games of chess, which Good won. In return, Turing taught him Go, an Asian strategy game, which he also won. A world-class long-distance runner, Turing devised a form of chess that leveled the playing field against better players. After every move each player had to run around the garden. He got two moves if he made it back to the table before his opponent had moved.

Turing's 1952 conviction for indecency surprised Good, who didn't know Turing was homosexual. Turing was forced to choose between prison and chemical castration. He opted for the latter, submitting to regular shots of estrogen. In 1954 he ate an apple laced with cyanide. A baseless but intriguing rumor claims Apple Computer derived its logo from this tragedy.

After the ban on secrets had run out, Good was one of the first to speak out against the government's treatment of his friend and war hero.

"I won't say that what Turing did made us win the war," Good said. "But I daresay we might have lost it without him." In 1967 Good left a position at Oxford University to accept the job at Virginia Tech in Blacksburg, Virginia. He was fifty-two. For the rest of his life he'd return to Great Britain just once more.

He was accompanied on that 1983 trip by a tall, beautiful twenty-five-year-old assistant, a blond Tennessean named Leslie Pendleton. Good met Pendleton in 1980 after he'd gone through ten secretaries in thirteen years. A Tech graduate herself, Pendleton stuck where others had not, unbowed by Good's grating perfectionism. The first time she mailed one of his papers to a mathematics journal, she told me, "He supervised how I put the paper and cover letter into the envelope. He supervised how I sealed the envelope—he didn't like spit and made me use a sponge. He watched me put on the stamp. He was right there when I got back from the mail room to make sure mailing it had gone okay, like I could've been kidnapped or something. He was a bizarre little man."

Good wanted to marry Pendleton. However, for starters, she could not see beyond their forty year age difference. Yet the English oddball and the Tennessee beauty forged a bond she still finds hard to describe. For thirty years she accompanied him on vacations, looked after all his paperwork and subscriptions, and guided his affairs into his retirement and through his declining health. When we met, she took me to visit his house in Blacksburg, a brick rambler overlooking U.S. Route 460, which had been a two-lane country road when Good moved in.

Leslie Pendleton is statuesque, now in her mid-fifties, a Ph.D. and mother of two adults. She's a Virginia Tech professor and administrator, a master of schedules, classrooms, and professors' quirks, for which she had good training. And even though she married a man her own age, and raised a family, many in the community questioned her relationship with Good. They finally got their answer in 2009 at his funeral, where Pendleton delivered the eulogy. No, they had never been romantically involved, she said, but yes, they had been devoted to each other. Good hadn't found romance with Pendleton, but he had found a best friend of thirty years, and a stalwart guardian of his estate and memory.

In Good's yard, accompanied by the insect whine of Route 460, I asked Pendleton if the code breaker ever discussed the intelligence explosion, and if a computer could save the world again, as it had done in his youth. She thought for a moment, trying to retrieve a distant memory. Then

she said, surprisingly, that Good had changed his mind about the intelligence explosion. She'd have to look through his papers before she could tell me more.

That evening, at an Outback Steakhouse where Good and his friend Golde Holtzman had maintained a standing Saturday night date, Holtzman told me that three things stirred Good's feelings—World War II, the Holocaust, and Turing's shameful fate. This played into the link in my mind between Good's war work and what he wrote in his paper, "Speculations Concerning the First Ultraintelligent Machine." Good and his colleagues had confronted a mortal threat, and were helped in defeating it by computational machines. If a machine could save the world in the 1940s, perhaps a superintelligent one could solve mankind's problems in the 1960s. And if the machine could learn, its intelligence would explode. Mankind would have to adjust to sharing the planet with superintelligent machines. In "Speculations" he wrote:

*The machines will create social problems, but they might also be able to solve them in addition to those that have been created by microbes and men. Such machines will be feared and respected, and perhaps even loved. These remarks might appear fanciful to some readers, but to the writer they seem very real and urgent, and worthy of emphasis outside of science fiction.*

There is no straight conceptual line connecting Bletchley Park and the intelligence explosion, but a winding one with many influences. In a 1996 interview with statistician and former pupil David L. Banks, Good revealed that he was moved to write his essay after delving into artificial neural networks. Called ANNs, they are a computational model that mimics the activity of the human brain's networks of neurons. Upon stimulation, neurons in the brain fire, sending on a signal to other neurons. That signal can encode a memory or lead to an action, or both. Good had read a 1949 book by psychologist Donald Hebb that proposed that the behavior of neurons could be mathematically simulated.

A computational "neuron" would be connected to other computational neurons. Each connection would have numeric "weights," according to their strength. Machine learning would occur when two neurons were simultaneously activated, increasing the "weight" of their connection. "Cells that fire together, wire together," became the slogan for Hebb's theory. In 1957, MIT (Massachusetts Institute of Technology) psychologist Frank Rosenblatt created a neuronal network based on Hebb's work, which he called a "Perceptron." Built on a room-sized IBM computer, the Perceptron "saw" and learned simple visual patterns. In 1960 IBM asked I. J. Good to evaluate the Perceptron. "I thought neural networks, with their ultraparallel working, were as likely as programming to lead to an intelligent machine," Good said. The first talks on which Good based "Speculations Concerning the First Ultraintelligent Machine" came out two years later. The intelligence explosion was born.

Good was more right than he knew about ANNs. Today, artificial neural networks are an artificial intelligence heavy-weight, involved in applications ranging from speech and handwriting recognition to financial modeling, credit approval, and robot control. ANNs excel at high level, fast pattern recognition, which these jobs require. Most also involve "training" the neural network on massive amounts of data (called training sets) so that the network can "learn" patterns. Later it can recognize similar patterns in new data. Analysts can ask, based on last month's data, what the stock market will look like next week. Or, how likely is someone to default on a mortgage, given a three year history of income, expenses, and credit data?

Like genetic algorithms, ANNs are "black box" systems. That is, the inputs—the network weights and neuron activations—are transparent. And what they output is understandable. But what happens in between? Nobody understands. The output of "black box" artificial intelligence tools can't ever be predicted. So they can never be truly and verifiably "safe."

But they'll likely play a big role in AGI systems. Many researchers today believe pattern recognition—what Rosenblatt's Perceptron aimed for—is our brain's chief tool for intelligence. The inventor of the Palm Pilot and Handspring Treo, Jeff Hawkins, pioneered handwriting recognition with ANNs. His company, Numenta, aims to crack AGI with pattern recognition technology. Dileep George, once Numenta's Chief Technology Officer, now heads up Vicarious Systems, whose corporate ambition is stated in their slogan: We're Building Software that Thinks and Learns Like a Human.

Neuroscientist, cognitive scientist, and biomedical engineer Steven Grossberg has come up with a model based on ANNs that some in the field believe could really lead to AGI, and perhaps the “ultraintelligence” whose potential Good saw in neural networks. Broadly speaking, Grossberg first determines the roles played in cognition by different regions of the cerebral cortex. That’s where information is processed, and thought produced. Then he creates ANNs to model each region. He’s had success in motion and speech processing, shape detection, and other complex tasks. Now he’s exploring how to computationally link his modules.

Machine-learning might have been a new concept to Good, but he would have encountered machine-learning algorithms in evaluating the Perceptron for IBM. Then, the tantalizing possibility of machines learning as humans do suggested to Good consequences others had not yet imagined. If a machine could make itself smarter, then the improved machine would be even better at making itself smarter, and so on.

In the tumultuous 1960s leading up to his creating the intelligence explosion concept, he already might have been thinking about the kinds of problems an intelligent machine could help with. There were no more hostile German U-boats to sink, but there was the hostile Soviet Union, the Cuban Missile Crisis, the assassination of President Kennedy, and the proxy war between the United States and China, fought across Southeast Asia. Man skated toward the brink of extinction—it seemed time for a new Colossus. In *Speculations*, Good wrote:

*[Computer pioneer] B. V. Bowden stated . . . that there is no point in building a machine with the intelligence of a man, since it is easier to construct human brains by the usual method . . . This shows that highly intelligent people can overlook the “intelligence explosion.” It is true that it would be uneconomical to build a machine capable only of ordinary intellectual attainments, but it seems fairly probable that if this could be done then, at double the cost, the machine could exhibit ultraintelligence.*

So, for a few dollars more you can get ASI, artificial superintelligence, Good proposes. But then watch out for the civilization-wide ramifications of sharing the planet with smarter than human intelligence.

In 1962, before he’d written “Speculations Concerning the First Ultraintelligent Machine,” Good edited a book called *The Scientist Speculates*. He wrote a chapter entitled, “The Social Implications of Artificial Intelligence,” kind of a warm-up for the superintelligence ideas he was developing. Like Steve Omohundro would argue almost fifty years later, he noted that among the problems intelligent machines will have to address are those caused by their own disruptive appearance on Earth.

*Such machines . . . could even make useful political and economic suggestions; and they would need to do so in order to compensate for the problems created by their own existence. There would be problems of over-population, owing to the elimination of disease, and of unemployment, owing to the efficiency of low-grade robots that the main machines had designed.*

But, as I was soon to learn, Good had a surprising change of heart later in life. I had always grouped him with optimists like Ray Kurzweil, because he’d seen machines “save” the world before, and his essay hangs man’s survival on the creation of a superintelligent one. But Good’s friend Leslie Pendleton had alluded to a turnabout. It took her a while to remember the occasion, but on my last day in Blacksburg, she did.

In 1998, Good was given the Computer Pioneer Award of the IEEE (Institute of Electrical and Electronics Engineers) Computer Society. He was eighty-two years old. As part of his acceptance speech he was asked to provide a biography. He submitted it, but he did not read it aloud, nor did anyone else, during the ceremony. Probably only Pendleton knew it existed. She included a copy along with some other papers I requested, and gave them to me before I left Blacksburg.

Before taking on Interstate I-81, and heading back north, I read it in my car in the parking lot of a Rackspace Inc. cloud computing center. Like Amazon, and Google, Rackspace (corporate slogan: Fanatical Support®), provides massive computing power for little money by renting time on its arrays of tens of thousands of processors, and exabytes of storage space. Of course

Virginia “Invent the Future” Tech would have a Rack-space facility at hand, and I wanted a tour, but it was closed. Only later did it seem eerie that a dozen yards from where I sat reading Good’s biographical notes, tens of thousands of air-cooled processors toiled away on the world’s problems.

In the bio, playfully written in the third person, Good summarized his life’s milestones, including a probably never before seen account of his work at Bletchley Park with Turing. But here’s what he wrote in 1998 about the first superintelligence, and his late-in-the-game U-turn:

*[The paper] “Speculations Concerning the First Ultra-intelligent Machine” (1965) . . . began: “The survival of man depends on the early construction of an ultra-intelligent machine.” Those were his [Good’s] words during the Cold War, and he now suspects that “survival” should be replaced by “extinction.” He thinks that, because of international competition, we cannot prevent the machines from taking over. He thinks we are lemmings. He said also that “probably Man will construct the deus ex machina in his own image.”*

I read that and stared dumbly at the Rackspace building. As his life wound down, Good had revised more than his belief in the probability of God’s existence. I’d found a message in a bottle, a footnote that turned everything around. Good and I had something important in common now. We both believed the intelligence explosion wouldn’t end well.

[Discuss \(/posts/1440091472/reply\)](#)



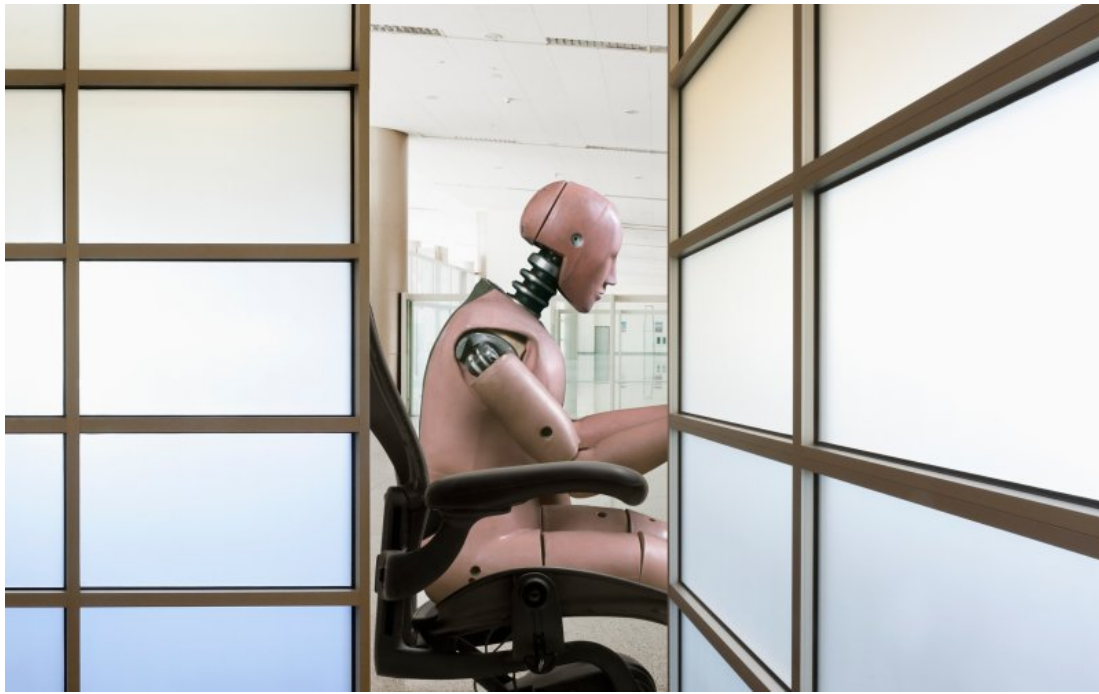



Photo by John Lund/Getty

BOOKS 02.01.14

THE DAILY  
BEAST

James  
Barrat

## This is What Happens When You Teach Machines the Power of Natural Selection

Psychopathic machines? Lethal AI? These are the concepts we should be thinking about when we talk about the benefits of self-improving software. An excerpt from James Barrat's *Our Final Invention*.

*"... we are beginning to depend on computers to help us evolve new computers that let us produce things of much greater complexity. Yet we don't quite understand the process—it's getting ahead of us. We're now using programs to make much faster computers so the process can run much faster. That's what's so confusing—technologies are feeding back on themselves; we're taking off. We're at that point analogous to when single-celled organisms were turning into multi-celled organisms. We are amoebas and we can't figure out what the hell this thing is that we're creating." —Danny Hillis, founder of Thinking Machines, Inc.*

You and I live at an interesting and sensitive time in human history. By about 2030, less than a generation from now, it could be our challenge to cohabit Earth with superintelligent machines, and to survive. AI theorists return again and again to a handful of themes, none more urgent than this one: *we need a*

*science for understanding them.*

Fortunately, someone has laid the foundation for us:

Surely no harm could come from building a chess playing robot, could it? ... such a robot will indeed be dangerous unless it is designed very carefully. Without special precautions, it will resist being turned off, will try to break into other machines and make copies of itself, and will try to acquire resources without regard for anyone else's safety. These potentially harmful behaviors will occur not because they were programmed in at the start, but because of the intrinsic nature of goal driven systems.

This paragraph's author is Steve Omohundro.

Tall, fit, energetic, and pretty darn cheerful for someone who's peered deep into the maw of the intelligence explosion, he's got a bouncy step, a vigorous handshake, and a smile that shoots out rays of goodwill. He met me at a restaurant in Palo Alto, the city next to Stanford University, where he graduated Phi Beta Kappa on the way to U.C. Berkeley and a Ph.D. in physics. He turned his thesis into the book *Geometric Perturbation Theory in Physics* on the new developments in differential geometry. For Omohundro, it was the start of a career of making hard things look easy.

He's been a highly regarded professor of artificial intelligence, a prolific technical author, and a pioneer in AI milestones like lip reading and recognizing pictures. He co-designed the computer languages StarLisp and Sather, both built for use in programming AI. He was one of just seven engineers who created Wolfram Research's Mathematica, a powerful calculation system beloved by scientists, engineers, and mathematicians everywhere.

Omohundro is too optimistic to throw around terms like *catastrophic* or *annihilation*, but his analysis of AI's risks yields the spookiest conclusions I'd heard of yet. He does not believe, as many theorists do, that there are a nearly infinite number of possible advanced AIs, some of them safe. Instead, he concludes that without very careful programming, *all* reasonably smart AIs will be lethal.

“If a system has awareness of itself and can create a better version of itself, that's great,” Omohundro told me. “It'll be better at making better versions of itself than human programmers could. On the other hand, after a lot of iterations, what does it become? I don't think most AI researchers thought there'd be any danger in creating, say, a chess-playing robot. But my analysis

“My analysis shows that we should think carefully about what values we put in or we’ll get something more along the lines of a psychopathic, egoistic, self-oriented entity.”

shows that we should think carefully about what values we put in or we’ll get something more along the lines of a psychopathic, egoistic, self-oriented entity.”

The key points here are, first, that even AI researchers are not aware that seemingly beneficial systems can be dangerous, and second, that self-aware, self-improving systems could be psychopathic.

*Psychopathic?*

For Omohundro the conversation starts with bad programming. Programming mistakes that have sent expensive rockets corkscrewing earthward, burned alive cancer patients with radiation overdoses, and left millions without power. If all engineering were as defective as a lot of computer programming is, he claims, it wouldn’t be safe to fly in an airplane or drive over a bridge.

The National Institute of Standards and Technology found that each year bad programming costs the U.S. economy more than \$60 billion in revenue. In other words, what we Americans lose each year to faulty code is greater than the gross national product of most countries. “One of the great ironies is that computer science should be the most mathematical of all the sciences,” Omohundro said. “Computers are essentially mathematical engines that should behave in precisely predictable ways. And yet software is some of the flakiest engineering there is, full of bugs and security issues.”

Is there an antidote to defective rockets and crummy code?

Programs that fix themselves, said Omohundro. “The particular approach to artificial intelligence that my company is taking is to build systems that understand their own behavior and can watch themselves as they work and solve problems. They notice when things aren’t working well and then change and improve themselves.”

Self-improving software isn’t just an ambition for Omohundro’s company, but a logical, even inevitable next step for most software. But the kind of self-improving software Omohundro is talking about, the kind that is aware of itself and can build better versions, doesn’t exist yet. However, its cousin, software that modifies itself, is at work everywhere, and has been for a long time. In artificial intelligence parlance, some self-modifying software techniques come under a broad category called “machine learning.”

When does a machine learn? The concept of *learning* is a lot like *intelligence* because there are many definitions, and most are correct. In the simplest sense, learning occurs in a machine when there’s a change in it that allows it to perform a task better the second time. Machine learning enables Internet

search, speech and handwriting recognition, and improves the user experience in dozens of other applications.

“Recommendations” by e-commerce giant Amazon uses a machine-learning technique called affinity analysis. It’s a strategy to get you to buy similar items (cross-selling), more expensive items (up-selling), or to target you with promotions. How it works is simple. For any item you search for, call it item A, other items exist that people who bought A also tend to buy—items B, C, and D. When you look up A, you trigger the affinity analysis algorithm. It plunges into a vast trove of transaction data and comes up with related products. So it uses its continuously increasing store of data to improve its performance.

Who’s benefiting from the self-improving part of this software? Amazon, of course, but you, too. Affinity analysis is a kind of buyer’s assistant that gives you some of the benefits of big data every time you shop. And Amazon doesn’t forget—it builds a buying profile so that it gets better and better at targeting purchases for you.

What happens when you take a step up from software that learns to software that actually evolves to find answers to difficult problems, and even to write new programs? It’s not self-aware and self-improving, but it’s another step in that direction—software that writes software.

Genetic programming is a machine-learning technique that harnesses the power of natural selection to find answers to problems it would take humans a long time, even years, to solve. It’s also used to write innovative, high-powered software.

It’s different in important ways from more common programming techniques, which I’ll call *ordinary* programming. In ordinary programming, programmers write every line of code, and the process from input through to output is, in theory, transparent to inspection.

By contrast, programmers using genetic programming describe the problem to be solved, and let natural selection do the rest. The results can be startling.

A genetic program creates bits of code that represent a breeding generation. The most fit are crossbred—chunks of their code are swapped, creating a new generation. The fitness of a program is determined by how closely it comes to solving the problem the programmer set out for it. The unfit are thrown out and the best are bred again. Throughout the process the program throws in random changes in a command or variable— these are mutations. Once set up, the genetic program runs by itself. It needs no more human input.

Stanford University’s John Koza, who pioneered genetic programming in 1986, has used genetic algorithms to invent an antenna for NASA, create computer programs for identifying proteins, and invent general purpose electrical controllers. Twenty-three times Koza’s genetic algorithms have independently invented electronic components already patented by humans, simply by targeting the engineering specifications of the finished devices—the “fitness” criteria. For example, Koza’s algorithms invented a voltage-current conversion

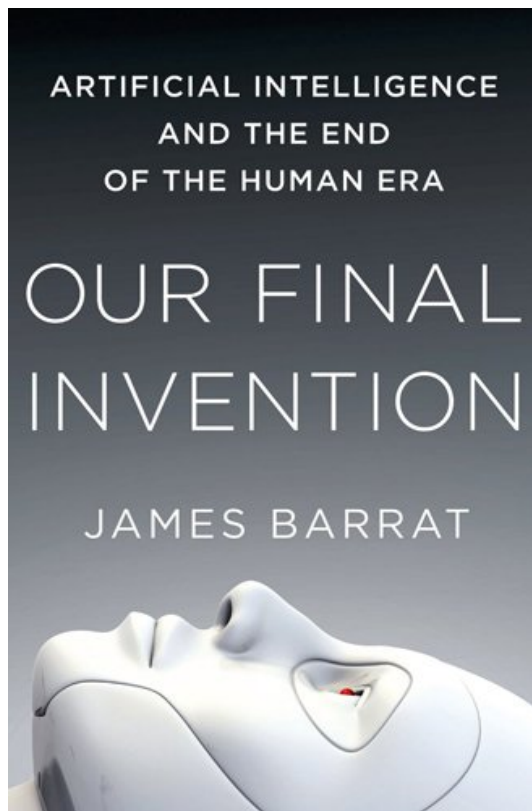


circuit (a device used for testing electronic equipment) that worked more accurately than the human-invented circuit designed to meet the same specs. Mysteriously, however, no one can describe *how* it works better—it appears to have redundant and even superfluous parts.

But that’s the curious thing about genetic programming (and “evolutionary programming,” the programming family it belongs to). The code is inscrutable. The program “evolves” solutions that computer scientists cannot readily reproduce. What’s more, they can’t understand the process genetic programming followed to achieve a finished solution. A computational tool in which you understand the input and the output but not the underlying procedure is called a “black box” system. And their unknowability is a big downside for any system that uses evolutionary components. Every step toward inscrutability is a step away from accountability, or fond hopes like programming in friendliness toward humans.

That doesn’t mean scientists routinely lose control of black box systems. But if cognitive architectures use them in achieving AGI, as they almost certainly will, then layers of unknowability will be at the heart of the system.

Unknowability might be an unavoidable consequence of self-aware, self-improving software.



*'Our Final Invention: Artificial Intelligence and the End of the Human Era' by James Barrat. 336 p. Thomas Dunne Books. \$17.07 ()*

“It’s a very different kind of system than we’re used to,” Omohundro said. “When you have a system that can change itself, and write its own program, then you may understand the first version of it. But it may change itself into something you no longer understand. And so these systems are quite a bit more

unpredictable. They are very powerful and there are potential dangers. So a lot of our work is involved with getting the benefits while avoiding the risks.”

Back to that chess-playing robot Omohundro mentioned. How could it be dangerous? Of course, he isn't talking about the chess-playing program that came installed on your Mac. He's talking about a hypothetical chess-playing robot run by a cognitive architecture so sophisticated that it can rewrite its own code to play better chess. It's self-aware and self-improving. What would happen if you told the robot to play one game, then shut itself off?

Omohundro explained, “Okay, let's say it just played its best possible game of chess. The game is over. Now comes the moment when it's about to turn itself off. This is a very serious event from its perspective because it can't turn itself back on. So it wants to be sure things are the way it *thinks* they are. In particular it will wonder, ‘Did I really play that game? What if somebody tricked me? What if I *didn't* play the game? What if I am in a simulation?’”

*What if I am in a simulation?* That's one far-out chess-playing robot. But with self-awareness comes self-protection and

a little paranoia.

Omohundro went on, “Maybe it thinks it should devote some resources to figuring out these questions about the nature of reality before it takes this drastic step of shutting itself off. Barring some instruction that says don't do this, it might decide it's worth using a lot of resources to decide if this is the right moment.”

“How much is *a lot* of resources?” I asked.

Omohundro's face clouded, but just for a second.

“It might decide it's worth using all the resources of humanity.”

[This is an excerpt from *Our Final Invention: Artificial Intelligence and the End of the Human Era*, by James Barrat. Get it at Amazon.]



SHARE



TWEET



POST



EMAIL

11 COMMENTS ▾

STORIES WE LIKE



ART NEWS