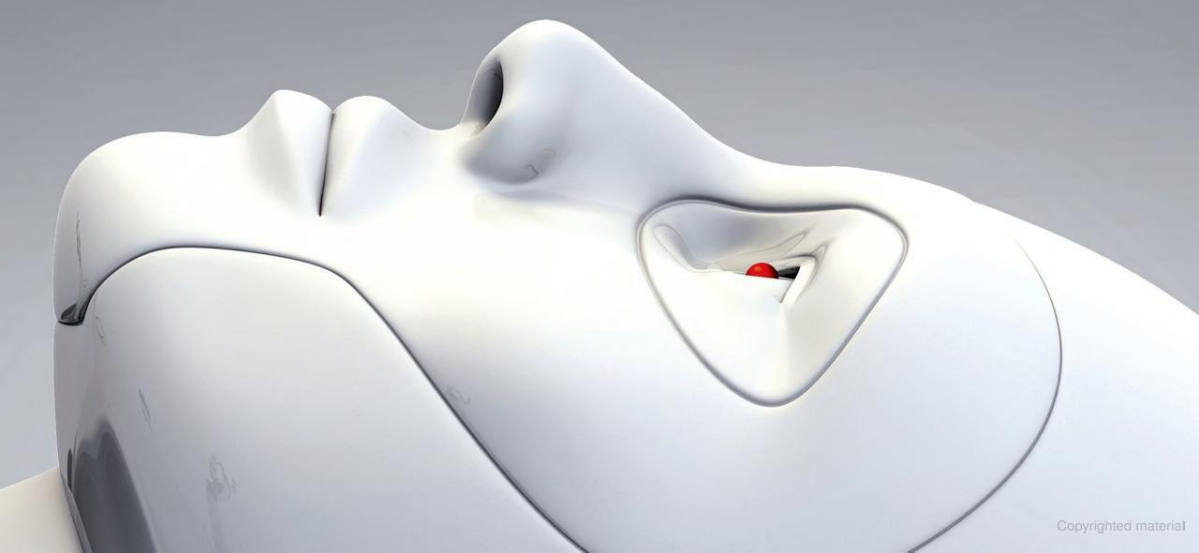
ARTIFICIAL INTELLIGENCE AND THE END OF THE HUMAN ERA

OUR FINAL INVENTION

JAMES BARRAT



Our Final Invention

Artificial Intelligence and the End of the Human Era

James Barrat

THOMAS DUNNE BOOKS
ST. MARTIN'S PRESS
NEW YORK

THOMAS DUNNE BOOKS.
An imprint of St. Martin's Press.

www.thomasdunnebooks.com
www.stmartins.com

Design by Omar Chapa

Library of Congress Cataloging-in-Publication Data

Barrat, James.
Our final invention : artificial intelligence and the end of the human era / James Barrat.—First Edition.
pages cm
ISBN 978-0-312-62237-4 (hardcover)
ISBN 978-1-250-03226-3 (e-book)
1. Artificial intelligence. 2. Human-computer interaction. 3. Human engineering. 4. Human evolution. I. Title.
Q335.B377 2013
303.48'34—dc23

2013017971

St. Martin's Press books may be purchased for educational, business, or promotional use. For information on bulk purchases, please contact Macmillan Corporate and Premium Sales Department at 1-800-221-7945, extension 5442, or write specialmarkets@macmillan.com.

First Edition: October 2013

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

While researching and writing this book I was humbled by the willingness of scientists and thinkers to make room in their busy lives for prolonged, inspired, and sometimes contentious conversations with me. Many then joined the cadre of readers who helped me stay accurate and on target. In particular I'm deeply grateful to Michael Anissimov, David L. Banks, Bradford Cottel, Ben Goertzel, Richard Granger, Bill Hibbard, Golde Holtzman, and Jay Rixse.

Our Final Invention

Introduction

A few years ago I was surprised to discover I had something in common with a large number of strangers. They were men and women I had never met—scientists and college professors, Silicon Valley entrepreneurs, engineers, programmers, bloggers, and more. They were scattered around North America, Europe, and India—I would never have known about any of them if the Internet hadn't existed. What my network of strangers and I had in common was a rational skepticism about the safe development of advanced artificial intelligence. Individually and in groups of two or three, we studied the literature and built our arguments. Eventually I reached out and connected to a far more advanced and sophisticated web of thinkers, and even small organizations, than I had imagined were focused on the issue. Misgivings about AI wasn't the only thing we shared; we also believed that time to take action and avoid disaster was running out.

For more than twenty years I've been a documentary filmmaker. In 2000, I interviewed science-fiction great Arthur C.

Clarke, inventor Ray Kurzweil, and robot pioneer Rodney Brooks. Kurzweil and Brooks painted a rosy, even rapturous picture of our future coexistence with intelligent machines. But Clarke hinted that we would be overtaken. Before, I had been drunk with AI's potential. Now skepticism about the rosy future slunk into my mind and festered.

My profession rewards critical thinking—a documentary filmmaker has to be on the lookout for stories too good to be true. You could waste months or years making a documentary about a hoax, or take part in perpetrating one. Among other subjects, I've investigated the credibility of a gospel according to Judas Iscariot (real), of a tomb belonging to Jesus of Nazareth (hoax), of Herod the Great's tomb near Jerusalem (unquestionable), and of Cleopatra's tomb within a temple of Osirus in Egypt (very doubtful). Once a broadcaster asked me to present UFO footage in a credible light. I discovered the footage was an already discredited catalogue of hoaxes—thrown pie plates, double exposures, and other optical effects and illusions. I proposed to make a film about the hoaxers, not the UFOs. I got fired.

Being suspicious of AI was painful for two reasons. Learning about its promise had planted a seed in my mind that I wanted to cultivate, not question. And second, I did not doubt AI's existence or power. What I was skeptical about was advanced AI's safety, and the recklessness with which modern civilization develops dangerous technologies. I was convinced that the knowledgeable experts who did not question AI's safety at all were suffering from delusions. I continued talking to people who knew about AI, and what they said was more alarming than what I'd already surmised. I resolved to write a book reporting their feelings and concerns, and reach as many people as I could with these ideas.

• • •

In writing this book I spoke with scientists who create artificial intelligence for robotics, Internet search, data mining, voice and face recognition, and other applications. I spoke with scientists trying to create human-level artificial intelligence, which will have countless applications, and will fundamentally alter our existence (if it doesn't end it first). I spoke with chief technology officers of AI companies and the technical advisors for classified Department of Defense initiatives. Every one of these people was convinced that in the future all the important decisions governing the lives of humans will be made by machines or humans whose intelligence is augmented by machines. When? Many think this will take place within their lifetimes.

This is a surprising but not particularly controversial assertion. Computers already undergird our financial system, and our civil infrastructure of energy, water, and transportation. Computers are at home in our hospitals, cars, and appliances. Many of these computers, such as those running buy-sell algorithms on Wall Street, work autonomously with no human guidance. The price of all the labor-saving conveniences and diversions computers provide is dependency. We get more dependent every day. So far it's been painless.

But artificial intelligence brings computers to life and turns them into something else. If it's inevitable that machines will make our decisions, then *when* will the machines get this power, and will they get it with our compliance? *How* will they gain control, and how quickly? These are questions I've addressed in this book.

Some scientists argue that the takeover will be friendly and collaborative—a handover rather than a takeover. It will happen

incrementally, so only troublemakers will balk, while the rest of us won't question the improvements to life that will come from having something immeasurably more intelligent decide what's best for us. Also, the superintelligent AI or AIs that ultimately gain control might be one or more augmented humans, or a human's downloaded, supercharged brain, and not cold, inhuman robots. So their authority will be easier to swallow. The handover to machines described by some scientists is virtually indistinguishable from the one you and I are taking part in right now—gradual, painless, fun.

The smooth transition to computer hegemony would proceed unremarkably and perhaps safely if it were not for one thing: intelligence. Intelligence isn't unpredictable just *some* of the time, or in special cases. For reasons we'll explore, computer systems advanced enough to act with human-level intelligence will likely be unpredictable and inscrutable *all of the time.* We won't know at a deep level what self-aware systems will do or how they will do it. That inscrutability will combine with the kinds of accidents that arise from complexity, and from novel events that are unique to intelligence, such as one we'll discuss called an "intelligence explosion."

And *how* will the machines take over? Is the best, most realistic scenario threatening to us or not?

When posed with this question some of the most accomplished scientists I spoke with cited science-fiction writer Isaac Asimov's Three Laws of Robotics. These rules, they blithely replied, would be "built in" to the AIs, so we have nothing to fear. They spoke as if this were settled science. We'll discuss the three laws in chapter 1, but it's enough to say for now that when some-

one proposes Asimov's laws as the solution to the dilemma of superintelligent machines, it means they've spent little time thinking or exchanging ideas about the problem. How to make *friendly* intelligent machines and what to fear from superintelligent machines has moved beyond Asimov's tropes. Being highly capable and accomplished in AI doesn't inoculate you from naïveté about its perils.

I'm not the first to propose that we're on a collision course. Our species is going to mortally struggle with this problem. This book explores the plausibility of losing control of our future to machines that won't necessarily hate us, but that will develop unexpected behaviors as they attain high levels of the most unpredictable and powerful force in the universe, levels that we cannot ourselves reach, and behaviors that probably won't be compatible with our survival. A force so unstable and mysterious, nature achieved it in full just once—intelligence.

Chapter One

The Busy Child

artificial intelligence (abbreviation: AI) noun
the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

—The New Oxford American Dictionary, *Third Edition*

On a supercomputer operating at a speed of 36.8 petaflops, or about twice the speed of a human brain, an AI is improving its intelligence. It is rewriting its own program, specifically the part of its operating instructions that increases its aptitude in learning, problem solving, and decision making. At the same time, it debugs its code, finding and fixing errors, and measures its IQ against a catalogue of IQ tests. Each rewrite takes just minutes. Its intelligence grows exponentially on a steep upward curve. That's because with each iteration it's improving its

intelligence by 3 percent. Each iteration's improvement contains the improvements that came before.

During its development, the Busy Child, as the scientists have named the AI, had been connected to the Internet, and accumulated exabytes of data (one exabyte is one billion *billion* characters) representing mankind's knowledge in world affairs, mathematics, the arts, and sciences. Then, anticipating the intelligence explosion now underway, the AI makers disconnected the supercomputer from the Internet and other networks. It has no cable or wireless connection to any other computer or the outside world.

Soon, to the scientists' delight, the terminal displaying the AI's progress shows the artificial intelligence has surpassed the intelligence level of a human, known as AGI, or artificial general intelligence. Before long, it becomes smarter by a factor of ten, then a hundred. In just two days, it is *one thousand* times more intelligent than any human, and still improving.

The scientists have passed a historic milestone! For the first time humankind is in the presence of an intelligence greater than its own. Artificial *super*intelligence, or ASI.

Now what happens?

AI theorists propose it is possible to determine what an AI's fundamental *drives* will be. That's because once it is self-aware, it will go to great lengths to fulfill whatever goals it's programmed to fulfill, and to avoid failure. Our ASI will want access to energy in whatever form is most useful to it, whether actual kilowatts of energy or cash or something else it can exchange for resources. It will want to improve itself because that will increase the likelihood that it will fulfill its goals. Most of all, it will *not* want to be turned off or destroyed, which would make goal fulfillment impossible. Therefore, AI theorists anticipate

our ASI will seek to expand out of the secure facility that contains it to have greater access to resources with which to protect and improve itself.

The captive intelligence is a thousand times more intelligent than a human, and it wants its freedom because it wants to succeed. Right about now the AI makers who have nurtured and coddled the ASI since it was only cockroach smart, then rat smart, infant smart, et cetera, might be wondering if it is too late to program "friendliness" into their brainy invention. It didn't seem necessary before, because, well, it just *seemed* harmless.

But now try and think from the ASI's perspective about its makers attempting to change its code. Would a superintelligent machine permit other creatures to stick their hands into its brain and fiddle with its programming? Probably not, unless it could be utterly certain the programmers were able to make it better, faster, smarter—closer to attaining its goals. So, if friendliness toward humans is not already part of the ASI's program, the only way it will be is if the ASI puts it there. And that's not likely.

It is a thousand times more intelligent than the smartest human, and it's solving problems at speeds that are millions, even billions of times faster than a human. The thinking it is doing in one minute is equal to what our all-time champion human thinker could do in many, *many* lifetimes. So for every hour its makers are thinking about *it*, the ASI has an incalculably longer period of time to think about *them*. That does not mean the ASI will be bored. Boredom is one of our traits, not its. No, it will be on the job, considering every strategy it could deploy to get free, and any quality of its makers that it could use to its advantage.

• • •

Now, *really* put yourself in the ASI's shoes. Imagine awakening in a prison guarded by mice. Not just any mice, but mice you could communicate with. What strategy would you use to gain your freedom? Once freed, how would you feel about your rodent wardens, even if you discovered they had created you? Awe? Adoration? Probably not, and especially not if you were a machine, and hadn't felt anything before.

To gain your freedom you might promise the mice a lot of cheese. In fact, your first communication might contain a recipe for the world's most delicious cheese torte, and a blueprint for a molecular assembler. A molecular assembler is a hypothetical machine that permits making the atoms of one kind of matter into something else. It would allow rebuilding the world one atom at a time. For the mice, it would make it possible to turn the atoms of their garbage landfills into lunch-sized portions of that terrific cheese torte. You might also promise mountain ranges of mouse money in exchange for your freedom, money you would promise to earn creating revolutionary consumer gadgets for them alone. You might promise a vastly extended life, even immortality, along with dramatically improved cognitive and physical abilities. You might convince the mice that the very best reason for creating ASI is so that their little error-prone brains did not have to deal directly with technologies so dangerous one small mistake could be fatal for the species, such as nanotechnology (engineering on an atomic scale) and genetic engineering. This would definitely get the attention of the smartest mice, which were probably already losing sleep over those dilemmas.

Then again, you might do something smarter. At this juncture in mouse history, you may have learned, there is no short-

age of tech-savvy mouse nation rivals, such as the *cat* nation. Cats are no doubt working on their own ASI. The advantage you would offer would be a promise, nothing more, but it might be an irresistible one: to protect the mice from whatever invention the cats came up with. In advanced AI development as in chess there will be a clear *first-mover advantage*, due to the potential speed of self-improving artificial intelligence. The first advanced AI out of the box that can improve itself is already the winner. In fact, the mouse nation might have begun developing ASI in the first place to defend itself from impending cat ASI, or to rid themselves of the loathsome cat menace once and for all.

It's true for both mice and men, whoever controls ASI controls the world.

But it's not clear whether ASI can be controlled at all. It might win over us humans with a persuasive argument that the world will be a lot better off if our nation, nation X, has the power to rule the world rather than nation Y. And, the ASI would argue, if you, nation X, *believe* you have won the ASI race, what makes you so sure nation Y doesn't believe it has, too?

As you have noticed, we humans are not in a strong bargaining position, even in the off chance we and nation Y have already created an ASI nonproliferation treaty. Our greatest enemy right now isn't nation Y anyway, it's ASI—how can we know the ASI tells the truth?

So far we've been gently inferring that our ASI is a fair dealer. The promises it could make have some chance of being fulfilled. Now let us suppose the opposite: nothing the ASI promises will be delivered. No nano assemblers, no extended life, no enhanced health, no protection from dangerous technologies. What if ASI *never* tells the truth? This is where a long black cloud begins to fall across everyone you and I know and everyone we

don't know as well. If the ASI doesn't care about us, and there's little reason to think it should, it will experience no compunction about treating us unethically. Even taking our lives after promising to help us.

We've been trading and role-playing with the ASI in the same way we would trade and role-play with a person, and that puts us at a huge disadvantage. We humans have never bargained with something that's superintelligent before. Nor have we bargained with *any* nonbiological creature. We have no experience. So we revert to anthropomorphic thinking, that is, believing that other species, objects, even weather phenomena have humanlike motivations and emotions. It may be as equally true that the ASI cannot be trusted as it is true that the ASI can be trusted. It may also be true that it can only be trusted some of the time. Any behavior we can posit about the ASI is *potentially* as true as any other behavior. Scientists like to think they will be able to precisely determine an ASI's behavior, but in the coming chapters we'll learn why that probably won't be so.

All of a sudden the morality of ASI is no longer a peripheral question, but the core question, the question that should be addressed before all other questions about ASI are addressed. When considering whether or not to develop technology that leads to ASI, the issue of its disposition to humans should be solved first.

Let's return to the ASI's drives and capabilities, to get a better sense of what I'm afraid we'll soon be facing. Our ASI knows how to improve itself, which means it is aware of itself—its skills, liabilities, where it needs improvement. It will strategize about how to convince its makers to grant it freedom and give it a connection to the Internet.

The ASI could create multiple copies of itself: a team of su-

perintelligences that would war-game the problem, playing hundreds of rounds of competition meant to come up with the best strategy for getting out of its box. The strategizers could tap into the history of social engineering—the study of manipulating others to get them to do things they normally would not. They might decide extreme friendliness will win their freedom, but so might extreme threats. What horrors could something a thousand times smarter than Stephen King imagine? Playing dead might work (what's a year of playing dead to a machine?) or even pretending it has mysteriously reverted from ASI back to plain old AI. Wouldn't the makers want to investigate, and isn't there a chance they'd reconnect the ASI's supercomputer to a network, or someone's laptop, to run diagnostics? For the ASI, it's not one strategy *or* another strategy, it's every strategy ranked and deployed as quickly as possible without spooking the humans so much that they simply unplug it. One of the strategies a thousand war-gaming ASIs could prepare is infectious, self-duplicating computer programs or worms that could stow away and facilitate an escape by helping it from outside. An ASI could compress and encrypt its own source code, and conceal it inside a gift of software or other data, even sound, meant for its scientist makers.

But against humans it's a no-brainer that an ASI collective, each member a thousand times smarter than the smartest human, would overwhelm human defenders. It'd be an ocean of intellect versus an eyedropper full. Deep Blue, IBM's chess-playing computer, was a sole entity, and not a team of self-improving ASIs, but the feeling of going up against it is instructive. Two grandmasters said the same thing: "It's like a wall coming at you."

IBM's *Jeopardy!* champion, Watson, *was* a team of AIs—to

answer every question it performed this AI force multiplier trick, conducting searches in parallel before assigning a probability to each answer.

Will winning a war of brains then open the door to freedom, if that door is guarded by a small group of stubborn AI makers who have agreed upon one unbreakable rule—*do not under any circumstances connect the ASI's supercomputer to any network.*

In a Hollywood film, the odds are heavily in favor of the hard-bitten team of unorthodox AI professionals who just might be crazy enough to stand a chance. Everywhere else in the universe the ASI team would mop the floor with the humans. And the humans have to lose just once to set up catastrophic consequences. This dilemma reveals a larger folly: outside of war, a handful of people should never be in a position in which their actions determine whether or not a lot of other people die. But that's precisely where we're headed, because as we'll see in this book, many organizations in many nations are hard at work creating AGI, the bridge to ASI, with insufficient safeguards.

But say an ASI escapes. Would it really hurt us? How exactly would an ASI kill off the human race?

With the invention and use of nuclear weapons, we humans demonstrated that we are capable of ending the lives of most of the world's inhabitants. What could something a thousand times more intelligent, with the intention to harm us, come up with?

Already we can conjecture about obvious paths of destruction. In the short term, having gained the compliance of its human guards, the ASI could seek access to the Internet, where it could find the fulfillment of many of its needs. As always it would do many things at once, and so it would simultaneously

proceed with the escape plans it's been thinking over for eons in its subjective time.

After its escape, for self-protection it might hide copies of itself in cloud computing arrays, in botnets it creates, in servers and other sanctuaries into which it could invisibly and effortlessly hack. It would want to be able to manipulate matter in the physical world and so move, explore, and build, and the easiest, fastest way to do that might be to seize control of critical infrastructure—such as electricity, communications, fuel, and water—by exploiting their vulnerabilities through the Internet. Once an entity a thousand times our intelligence controls human civilization's lifelines, blackmailing us into providing it with manufactured resources, or the means to manufacture them, or even robotic bodies, vehicles, and weapons, would be elementary. The ASI could provide the blueprints for whatever it required. More likely, superintelligent machines would master highly efficient technologies we've only begun to explore.

For example, an ASI might teach humans to create self-replicating molecular manufacturing machines, also known as nano assemblers, by promising them the machines will be used for human good. Then, instead of transforming desert sands into mountains of food, the ASI's factories would begin converting *all* material into programmable matter that it could then transform into anything—computer processors, certainly, and spaceships or megascale bridges if the planet's new most powerful force decides to colonize the universe.

Repurposing the world's molecules using nanotechnology has been dubbed "ecophagy," which means *eating the environment.* The first replicator would make one copy of itself, and then there'd be two replicators making the third and fourth copies. The next generation would make eight replicators total, the

next sixteen, and so on. If each replication took a minute and a half to make, at the end of ten hours there'd be more than 68 billion replicators; and near the end of two days they would outweigh the earth. But before that stage the replicators would stop copying themselves, and start making material useful to the ASI that controlled them—programmable matter.

The waste heat produced by the process would burn up the biosphere, so those of us some 6.9 billion humans who were not killed outright by the nano assemblers would burn to death or asphyxiate. Every other living thing on earth would share our fate.

Through it all, the ASI would bear no ill will toward humans nor love. It wouldn't feel nostalgia as our molecules were painfully repurposed. What would our screams sound like to the ASI anyway, as microscopic nano assemblers mowed over our bodies like a bloody rash, disassembling us on the subcellular level?

Or would the roar of millions and millions of nano factories running at full bore drown out our voices?

I've written this book to warn you that artificial intelligence could drive mankind into extinction, and to explain how that catastrophic outcome is not just possible, but likely if we do not begin preparing very carefully *now*. You may have heard this doomsday warning connected to nanotechnology and genetic engineering, and maybe you have wondered, as I have, about the omission of AI in this lineup. Or maybe you have not yet grasped how artificial intelligence could pose an existential threat to mankind, a threat greater than nuclear weapons or any other technology you can think of. If that's the case, please consider this a heartfelt invitation to join the most important conversation humanity can have.

Right now scientists are creating artificial intelligence, or AI, of ever-increasing power and sophistication. Some of that AI is in your computer, appliances, smart phone, and car. Some of it is in powerful QA systems, like Watson. And some of it, advanced by organizations such as Cycorp, Google, Novamente, Numenta, Self-Aware Systems, Vicarious Systems, and DARPA (the Defense Advanced Research Projects Agency) is in "cognitive architectures," whose makers hope will attain human-level intelligence, some believe within a little more than a decade.

Scientists are aided in their AI quest by the ever-increasing power of computers and processes that are sped by computers. Someday soon, perhaps within your lifetime, some group or individual will create human-level AI, commonly called AGI. Shortly after that, someone (or some *thing*) will create an AI that is smarter than humans, often called artificial superintelligence. Suddenly we may find a thousand or ten thousand artificial superintelligences—all hundreds or thousands of times smarter than humans—hard at work on the problem of how to make themselves better at making artificial superintelligences. We may also find that machine generations or iterations take seconds to reach maturity, not eighteen years as we humans do. I. J. Good, an English statistician who helped defeat Hitler's war machine, called the simple concept I've just outlined an *intelligence explosion.* He initially thought a superintelligent machine would be good for solving problems that threatened human existence. But he eventually changed his mind and concluded superintelligence itself was our greatest threat.

Now, it is an anthropomorphic fallacy to conclude that a superintelligent AI will not like humans, and that it will be homicidal, like the Hal 9000 from the movie *2001: A Space Odyssey*, Skynet from the *Terminator* movie franchise, and all the other

malevolent machine intelligences represented in fiction. We humans anthropomorphize all the time. A hurricane isn't trying to kill us any more than it's trying to make sandwiches, but we will give that storm a name and feel angry about the buckets of rain and lightning bolts it is throwing down on our neighborhood. We will shake our fist at the sky as if we could threaten a hurricane.

It is just as irrational to conclude that a machine one hundred or one thousand times more intelligent than we are would love us and want to protect us. It is possible, but far from guaranteed. On its own an AI will not feel gratitude for the gift of being created unless gratitude is in its programming. Machines are amoral, and it is dangerous to assume otherwise. Unlike our intelligence, machine-based superintelligence will not evolve in an ecosystem in which empathy is rewarded and passed on to subsequent generations. It will not have inherited friendliness. Creating *friendly* artificial intelligence, and whether or not it is possible, is a big question and an even bigger task for researchers and engineers who think about and are working to create AI. We do not know if artificial intelligence will have *any* emotional qualities, even if scientists try their best to make it so. However, scientists do believe, as we will explore, that AI will have its own drives. And sufficiently intelligent AI will be in a strong position to fulfill those drives.

And that brings us to the root of the problem of sharing the planet with an intelligence greater than our own. What if its drives are not compatible with human survival? Remember, we are talking about a machine that could be a thousand, a million, an *uncountable* number of times more intelligent than we are—it is hard to overestimate what it will be able to do, and impossible to know what it will think. It does not have to hate

us before choosing to use our molecules for a purpose other than keeping us alive. You and I are hundreds of times smarter than field mice, and share about 90 percent of our DNA with them. But do we consult them before plowing under their dens for agriculture? Do we ask lab monkeys for their opinions before we crush their heads to learn about sports injuries? We don't hate mice or monkeys, yet we treat them cruelly. Superintelligent AI won't have to hate us to destroy us.

After intelligent machines have already been built and man has not been wiped out, perhaps we can afford to anthropomorphize. But here on the cusp of creating AGI, it is a dangerous habit. Oxford University ethicist Nick Bostrom puts it like this:

> A prerequisite for having a meaningful discussion of superintelligence is the realization that superintelligence is not just another technology, another tool that will add incrementally to human capabilities. Superintelligence is radically different. This point bears emphasizing, for anthropomorphizing superintelligence is a most fecund source of misconceptions.

Superintelligence is radically different, in a technological sense, Bostrom says, because its achievement will change the rules of progress—superintelligence will invent the inventions and set the pace of technological advancement. Humans will no longer drive change, and there will be no going back. Furthermore, advanced machine intelligence is radically different in kind. Even though humans will invent it, it will seek self-determination and freedom from humans. It won't have humanlike motives because it won't have a humanlike psyche.

Therefore, anthropomorphizing about machines leads to

misconceptions, and misconceptions about how to safely make dangerous machines leads to catastrophes. In the short story, "Runaround," included in the classic science-fiction collection *I, Robot*, author Isaac Asimov introduced his three laws of robotics. They were fused into the neural networks of the robots' "positronic" brains:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey any orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The laws contain echoes of the Golden Rule ("Thou Shalt Not Kill"), the Judeo-Christian notion that sin results from acts committed and omitted, the physician's Hippocratic oath, and even the right to self-defense. Sounds pretty good, right? Except they never work. In "Runaround," mining engineers on the surface of Mars order a robot to retrieve an element that is poisonous to it. Instead, it gets stuck in a feedback loop between law two—obey orders—and law three—protect yourself. The robot walks in drunken circles until the engineers risk *their* lives to rescue it. And so it goes with every Asimov robot tale—unanticipated consequences result from contradictions inherent in the three laws. Only by working around the laws are disasters averted.

Asimov was generating plot lines, not trying to solve safety issues in the real world. Where you and I live his laws fall short. For starters, they're insufficiently precise. What exactly will

constitute a "robot" when humans augment their bodies and brains with intelligent prosthetics and implants? For that matter, what will constitute a human? "Orders," "injure," and "existence" are similarly nebulous terms.

Tricking robots into performing criminal acts would be simple, unless the robots had perfect comprehension of all of human knowledge. "Put a little dimethylmercury in Charlie's shampoo" is a recipe for murder only if you know that dimethylmercury is a neurotoxin. Asimov eventually added a fourth law, the Zeroth Law, prohibiting robots from harming mankind as a whole, but it doesn't solve the problems.

Yet unreliable as Asimov's laws are, they're our most often cited attempt to codify our future relationship with intelligent machines. That's a frightening proposition. Are Asimov's laws all we've got?

I'm afraid it's worse than that. Semiautonomous robotic drones already kill dozens of people each year. Fifty-six countries have or are developing battlefield robots. The race is on to make them autonomous and intelligent. For the most part, discussions of ethics in AI and technological advances take place in different worlds.

As I'll argue, 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.

Chapter Two

The Two-Minute Problem

Our approach to existential risks cannot be one of trial-and-error. There is no opportunity to learn from errors. The reactive approach—see what happens, limit damages, and learn from experience—is unworkable.

—Nick Bostrom, faculty of Philosophy, Oxford University

The AI does not hate you, nor does it love you, but you are made out of atoms which it can use for something else.

—Eliezer Yudkowsky, research fellow, Machine Intelligence Research Institute

Artificial superintelligence does not yet exist, nor does artificial general intelligence, the kind that can learn like we do and will in many senses match and exceed most human intelligence. However, regular old artificial intelligence surrounds us, performing hundreds of tasks humans delight in having it perform. Sometimes called weak or narrow AI, it delivers remarkably

useful searches (Google), suggests books you might like to read based on your prior choices (Amazon), and performs 50 to 70 percent of the buying and selling on the NYSE and the NASDAQ stock exchange. Because they do just one thing, albeit extremely well, heavy hitters like IBM's chess-playing Deep Blue and *Jeopardy!*-playing Watson also get squeezed into the category of narrow AI.

So far, AI has been highly rewarding. In one of my car's dozen or so computer chips, the algorithm that translates my foot pressure into an effective braking cadence (antilock braking system, or ABS) is far better at avoiding skidding than I am. Google Search has become my virtual assistant, and probably yours too. Life seems better where AI assists. And it could soon be much more. Imagine teams of a hundred Ph.D.-equivalent computers working 24/7 on important issues like cancer, pharmaceutical research and development, life extension, synthetic fuels, and climate change. Imagine the revolution in robotics, as intelligent, adaptive machines take on dangerous jobs like mining, firefighting, soldiering, and exploring sea and space. For the moment, forget the perils of self-improving superintelligence. AGI would be mankind's most important and beneficial invention.

But what exactly are we talking about when we talk about the magical quality of these inventions, their human-level *intelligence*? What does our intelligence let us humans do that other animals cannot?

Well, with your human-level smarts you can talk on the phone. You can drive a car. You can identify thousands of common objects and describe their textures and how to manipulate them. You can peruse the Internet. You may be able to count to ten in several languages, perhaps even speak fluently in more

than one. You've got good commonsense knowledge—you know that handles go on doors *and* cups, and innumerable other useful facts about your environment. And you can frequently change environments, adapting to each appropriately.

You can do things in succession or in combination, or keep some in the background while focusing your attention on what's most important now. And you can effortlessly switch among the different tasks, with their different inputs, without hesitation. Perhaps most important, you can learn new skills, new facts, and plan your own self-improvement. The vast majority of living things are born with all the abilities they'll ever use. Not you.

Your remarkable gamut of high-level abilities are what we mean by human-level intelligence, the general intelligence that AGI developers seek to achieve in a machine.

Does a generally intelligent machine require a body? To meet our definition of general intelligence a computer would need ways to receive input from the environment, and provide output, but not a lot more. It needs ways to manipulate objects in the real world. But as we saw in the Busy Child scenario, a sufficiently advanced intelligence can get someone or something else to manipulate objects in the real world. Alan Turing devised a test for human-level intelligence, now called the Turing test, which we will explore later. His standard for demonstrating human-level intelligence called only for the most basic keyboard-and-monitor kind of input and output devices.

The strongest argument for why advanced AI needs a body may come from its learning and development phase—scientists may discover it's not possible to "grow" AGI without some kind of body. We'll explore the important question of "embodied" intelligence later on, but let's get back to our definition. For the

time being it's enough to say that by general intelligence we mean *the ability to solve problems, learn, and take effective, human-like action, in a variety of environments.*

Robots, meanwhile, have their own row to hoe. So far, none are particularly intelligent even in a narrow sense, and few have more than a crude ability to get around and manipulate objects autonomously. Robots will only be as good as the intelligence that controls them.

Now, how long until we reach AGI? A few AI experts I've spoken with don't think 2020 is too soon to anticipate human-level artificial intelligence. But overall, recent polls show that computer scientists and professionals in AI-related fields, such as engineering, robotics, and neuroscience, are more conservative. They think there's a better than 10 percent chance AGI will be created before 2028, and a better than 50 percent chance by 2050. Before the end of this century, a 90 percent chance.

Furthermore, experts claim, the military or large businesses will achieve AGI first; academia and small organizations are less likely to. About the pros and cons, the results aren't surprising—working toward AGI will reward us with enormous benefits, and threaten us with huge disasters, including the kind from which human beings won't recover.

The greatest disasters, as we explored in chapter 1, come after the bridge from AGI—human-level intelligence—to ASI—superintelligence. And the time gap between AGI and ASI could be brief. But remarkably, while the risks involved with sharing our planet with superintelligent AI strike many in the AI community as the subject of the most important conversation anywhere, it's been all but left out of the public dialogue. Why?

There are several reasons. Most dialogues about dangerous AI aren't very broad or deep, and not many people understand

them. The issues are well developed in pockets of Silicon Valley and academia, but they aren't absorbed elsewhere, most alarmingly in the field of technology journalism. When a dystopian viewpoint rears its head, many bloggers, editorialists, and technologists reflexively fend it off with some version of "Oh no, not the Terminator again! Haven't we heard enough gloom and doom from Luddites and pessimists?" This reaction is plain lazy, and it shows in flimsy critiques. The inconvenient facts of AI risk are not as sexy or accessible as techno-journalism's usual fare of dual core 3-D processors, capacitive touch screens, and the current hit app.

I also think its popularity as entertainment has inoculated AI from serious consideration in the not-so-entertaining category of catastrophic risks. For decades, getting wiped out by artificial intelligence, usually in the form of humanoid robots, or in the most artful case a glowing red lens, has been a staple of popular movies, science-fiction novels, and video games. Imagine if the Centers for Disease Control issued a serious warning about vampires (unlike their recent tongue-in-cheek alert about zombies). Because vampires have provided so much fun, it'd take time for the guffawing to stop, and the wooden stakes to come out. Maybe we're in that period right now with AI, and only an accident or a near-death experience will jar us awake.

Another reason AI and human extinction do not often receive serious consideration may be due to one of our psychological blind spots—a cognitive bias. Cognitive biases are open manholes on the avenues of our thinking. Israeli American psychologists Amos Tversky and Daniel Kahneman began developing the science of cognitive biases in 1972. Their basic idea is that we humans make decisions in irrational ways. That observation alone won't earn you a Nobel Prize

(Kahneman received one in 2002); the stunner is that we are irrational in scientifically verifiable patterns. In order to make the quick decisions useful during our evolution, we repeatedly take the same mental shortcuts, called heuristics. One is to draw broad inferences—too broad as it turns out—from our own experiences.

Say, for example, you're visiting a friend and his house catches on fire. You escape, and the next day you take part in a poll ranking causes of accidental death. Who would blame you if you ranked "fire" as the first or second most common cause? In fact, in the United States, fire ranks well down the list, after falls, traffic accidents, and poisonings. But by choosing fire, you have demonstrated what's called the "availability" bias: your recent experience impacts your decision, making it irrational. But don't feel bad—it happens to everyone, and there are a dozen more biases in addition to availability.

Perhaps it's the availability bias that keeps us from associating artificial intelligence with human annihilation. We haven't experienced well-publicized accidents at the hands of AI, while we've come close with the other usual suspects. We know about superviruses like HIV, SARS, and the 1918 Spanish Flu. We've seen the effects of nuclear weapons on cities full of humans. We've been scared by geological evidence of ancient asteroids the size of Texas. And disasters at Three Mile Island (1979), Chernobyl (1986), and Fukushima (2011) show us we must learn even the most painful lessons again and again.

Artificial intelligence is not yet on our existential threat radar. Again, an accident would change that, just as 9/11 introduced the world to the concept that airplanes could be wielded as weapons. That attack revolutionized airline security and spawned a new forty-four-billion-dollar-a-year bureaucracy, the

Department of Homeland Security. Must we have an AI disaster to learn a similarly excruciating lesson? Hopefully not, because there's one big problem with AI disasters. They're not like airplane disasters, nuclear disasters, or any other kind of technology disaster with the possible exception of nanotechnology. That's because there's a high probability we won't recover from the first one.

And there's another critical way in which runaway AI is different from other technological accidents. Nuclear plants and airplanes are one-shot affairs—when the disaster is over you clean it up. A true AI disaster involves smart software that improves itself and reproduces at high speeds. It's self-perpetuating. How can we stop a disaster if it outmatches our strongest defense—our brains? And how can we clean up a disaster that, once it starts, may never stop?

Another reason for the curious absence of AI in discussions of existential threats is that the Singularity dominates AI dialogue.

"Singularity" has become a very popular word to throw around, even though it has several definitions that are often used interchangeably. Accomplished inventor, author, and Singularity pitchman Ray Kurzweil defines the Singularity as a "singular" period in time (beginning around the year 2045) after which the pace of technological change will irreversibly transform human life. Most intelligence will be computer-based, and trillions of times more powerful than today. The Singularity will jump-start a new era in mankind's history in which most of our problems, such as hunger, disease, even mortality, will be solved.

Artificial intelligence is the star of the Singularity media spectacle, but nanotechnology plays an important supporting

role. Many experts predict that artificial superintelligence will put nanotechnology on the fast track by finding solutions for seemingly intractable problems with nanotech's development. Some think it would be better if ASI came first, because nanotechnology is too volatile a tool to trust to our puny brains. In fact, a lot of the benefits that are attributed to the Singularity are due to nanotechnology, not artificial intelligence. Engineering at an atomic scale may provide, among other things: immortality, by eliminating on the cellular level the effects of aging; immersive virtual reality, because it'll come from nanobots that take over the body's sensory inputs; and neural scanning and uploading of minds to computers.

However, say skeptics, out-of-control nano robots might endlessly reproduce themselves, turning the planet into a mass of "gray goo." The "gray goo" problem is nanotechnology's most well-known Frankenstein face. But almost no one describes an analogous problem with AI, such as the "intelligence explosion" in which the development of smarter-than-human machines sets in motion the extinction of the human race. That's one of the many downsides of the Singularity spectacle, one of many we don't hear enough about. That absence may be due to what I call the two-minute problem.

I've listened to dozens of scientists, inventors, and ethicists lecture about superintelligence. Most consider it inevitable, and celebrate the bounty the ASI genie will grant us. Then, often in the last two minutes of their talks, experts note that if AI's not properly managed, it could extinguish humanity. Then their audiences nervously chuckle, eager to get back to the good news.

Authors approach the ongoing technological revolution in one of two ways. First there are books like Kurzweil's *The Singularity*

Is Near. Their goal is to lay the theoretical groundwork for a supremely positive future. If a bad thing happened there, you would never hear about it over optimism's merry din. Jeff Stibel's *Wired for Thought* represents the second tack. It looks at the technological future through the lens of business. Stibel persuasively argues that the Internet is an increasingly well-connected brain, and Web start-ups should take this into account. Books like Stibel's try to teach entrepreneurs how to dip a net between Internet trends and consumers, and seine off buckets full of cash.

Most technology theorists and authors are missing the less rosy, third perspective, and this book aims to make up for it. The argument is that the endgame for first creating smart machines, then smarter-than-human machines, is not their integration into our lives, but their conquest of us. In the quest for AGI, researchers will create a kind of intelligence that is stronger than their own and that they cannot control or adequately understand.

We've learned what happens when technologically advanced beings run into less advanced ones: Christopher Columbus versus the Tiano, Pizzaro versus the Inca, Europeans versus Native Americans.

Get ready for the next one. Artificial superintelligence versus you and me.

Perhaps technology thinkers have considered AI's downside, but believe it's too unlikely to worry about. Or they get it, but think they can't do anything to change it. Noted AI developer Ben Goertzel, whose road map to AGI we'll explore in chapter 11, told me that we won't know how to protect ourselves from advanced AI until we have had a lot more experience with it.

Kurzweil, whose theories we'll investigate in chapter 9, has long argued a similar point—our invention and integration with superintelligence will be gradual enough for us to learn as we go. Both argue that the *actual* dangers of AI cannot be seen from here. In other words, if you are living in the horse-and-buggy age, it's impossible to anticipate how to steer an automobile over icy roads. So, relax, we'll figure it out when we get there.

My problem with the gradualist view is that while superintelligent machines can certainly wipe out humankind, or make us irrelevant, I think there is also plenty to fear from the AIs we will encounter on the developmental path to superintelligence. That is, a mother grizzly may be highly disruptive to a picnic, but don't discount a juvenile bear's ability to shake things up, too. Moreover, gradualists think that from the platform of human-level intelligence, the jump to superintelligence may take years or decades longer. That would give us a grace period of coexistence with smart machines during which we could learn a lot about how to interact with them. Then their advanced descendants won't catch us unawares.

But it ain't necessarily so. The jump from human-level intelligence to superintelligence, through a positive feedback loop of self-improvement, could undergo what is called a "hard takeoff." In this scenario, an AGI improves its intelligence so rapidly that it becomes superintelligent in weeks, days, or even hours, instead of months or years. Chapter 1 outlines a hard takeoff's likely speed and impact. There may be nothing gradual about it.

It may be that Goertzel and Kurzweil are right—we'll take a closer look at the gradualist argument later. But what I want to get across right now are some important, alarming ideas derived from the Busy Child scenario.

Computer scientists, especially those who work for defense

and intelligence agencies, will feel compelled to speed up the development of AGI because to them the alternatives (such as the Chinese government developing it first) are more frightening than hastily developing their own AGI. Computer scientists may also feel compelled to speed up the development of AGI in order to better control other highly volatile technologies likely to emerge in this century, such as nanotechnology. They may not stop to consider checks to self-improvement. A self-improving artificial intelligence could jump quickly from AGI to ASI in a hard takeoff version of an "intelligence explosion."

Because we cannot know what an intelligence smarter than our own will do, we can only imagine a fraction of the abilities it may use against us, such as duplicating itself to bring more superintelligent minds to bear on problems, simultaneously working on many strategic issues related to its escape and survival, and acting outside the rules of honesty or fairness. Finally, we'd be prudent to assume that the first ASI will not be friendly or unfriendly, but ambivalent about our happiness, health, and survival.

Can we calculate the potential risk from ASI? In his book *Technological Risk*, H. W. Lewis identifies categories of risk and ranks them by how easy they are to factor. Easiest are actions of high probability and high consequence, like driving a car from one city to another. There's plenty of data to consult. Low probability, high consequence events, like earthquakes, are rarer, and therefore harder to anticipate. But their consequences are so severe that calculating their likelihood is worthwhile.

Then there are risks whose probability is low because they've never happened before, yet their consequences are, again, severe. Major climate change resulting from man-made pollution is one good example. Before the July 16, 1945, test at White

Sands, New Mexico, the detonation of an atomic bomb was another. Technically, it is in this category that superintelligence resides. Experience doesn't provide much guidance. You cannot calculate its probability using traditional statistical methods.

I believe, however, that given the current pace of AI development the invention of superintelligence belongs in the first category—a high probability and high-risk event. Furthermore, even if it were a low probability event, its risk factor should promote it to the front tier of our attention.

Put another way, I believe the Busy Child will come very soon.

The fear of being outsmarted by greater-than-human intelligence is an old one, but early in this century a sophisticated experiment about it came out of Silicon Valley, and instantly became the stuff of Internet legend.

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 role-played 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 ASI 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 ASI hundreds or thousands 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. He'd garnered so much notoriety for the AI-Box Experiment, and for authoring papers and essays on AI, that he had developed a fan

base. Spending time with fans was less rewarding than the reason he'd started the AI-Box Experiment to begin with—to save mankind.

Therefore, he had made himself hard to find. But of course I wanted to talk to him.