The AI Cargo Cult

The Myth of a Superhuman AI

Kevin Kelly

WIRED April 25, 2017

I've heard that in the future computerized AIs will become so much smarter than us that they will take all our jobs and resources, and humans will go extinct. Is this true?

That's the most common question I get whenever I give a talk about AI. The questioners are earnest; their worry stems in part from some experts who are asking themselves the same thing. These folks are some of the smartest people alive today, such as Stephen Hawking, Elon Musk, Max Tegmark, Sam Harris, and Bill Gates, and they believe this scenario very likely could be true. Recently at a conference convened to discuss these AI issues, a panel of nine of the most informed gurus on AI all agreed this superhuman intelligence was inevitable and not far away.

Yet buried in this scenario of a takeover of superhuman artificial intelligence are five assumptions which, when examined closely, are not based on any evidence. These claims might be true in the future, but there is no evidence to date to support them. The assumptions behind a superhuman intelligence arising soon are:

1. Artificial intelligence is already getting smarter than us, at an exponential rate.
2. We'll make AIs into a general purpose intelligence, like our own.
3. We can make human intelligence in silicon.
4. Intelligence can be expanded without limit.
5. Once we have exploding superintelligence it can solve most of our problems.

In contradistinction to this orthodoxy, I find the following five heresies to have more evidence to support them.

1. Intelligence is not a single dimension, so "smarter than humans" is a meaningless concept.
2. Humans do not have general purpose minds, and neither will AIs.
3. Emulation of human thinking in other media will be constrained by cost.
4. Dimensions of intelligence are not infinite.
5. Intelligences are only one factor in progress.

If the expectation of a superhuman AI takeover is built on five key assumptions that have no basis in evidence, then this idea is more akin to a religious belief — a myth. In the following paragraphs I expand my evidence for each of these five counter-assumptions, and make the case that, indeed, a superhuman AI is a kind of myth.

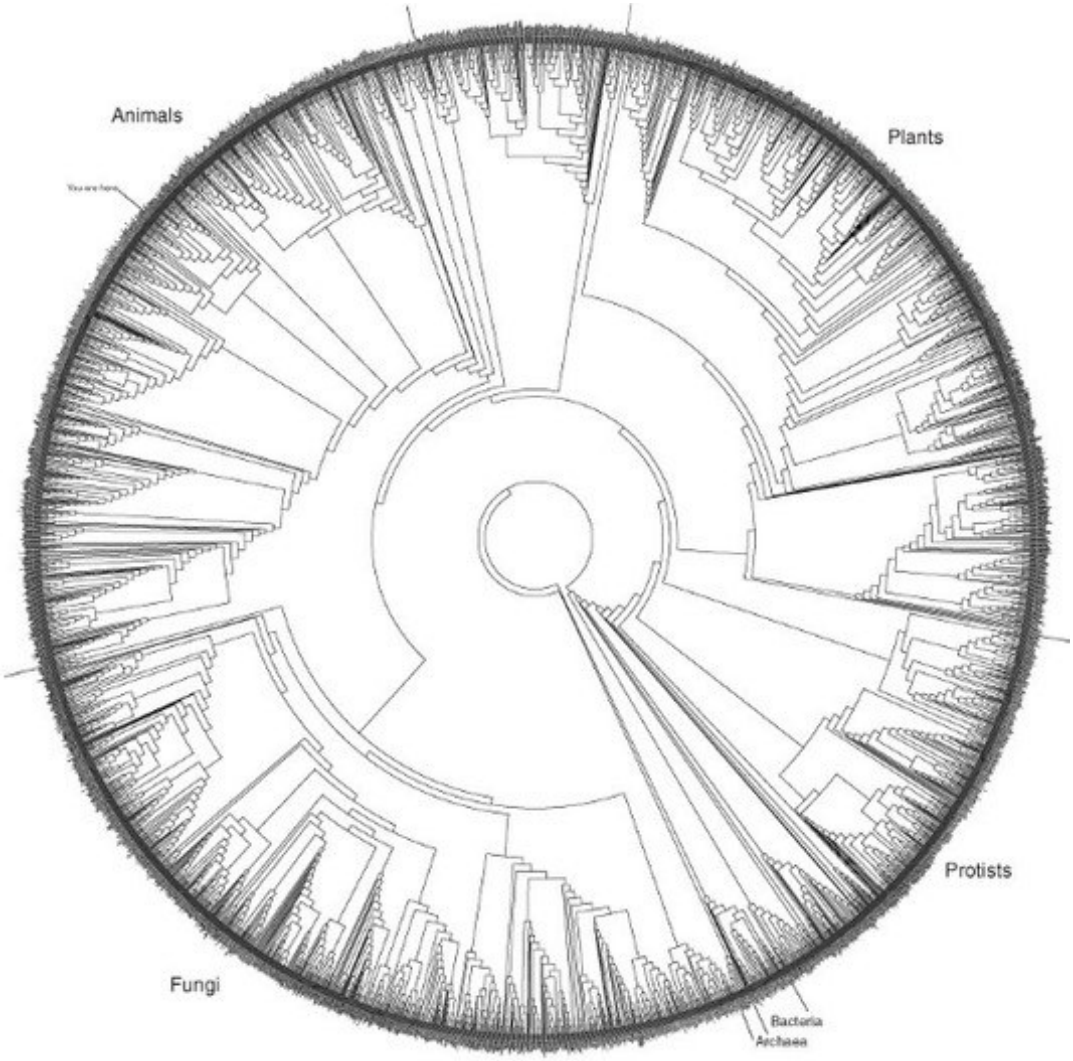
1.

The most common misconception about artificial intelligence begins with the common misconception about natural intelligence. This misconception is that intelligence is a single dimension. Most technical people tend to graph intelligence the way Nick Bostrom does in his book, Superintelligence — as a literal, single-dimension, linear graph of increasing amplitude. At one end is the low intelligence of, say, a small animal; at the other end is the high intelligence, of, say, a genius—almost as if intelligence were a sound level in decibels. Of course, it is then very easy to imagine the extension so that the loudness of intelligence continues to grow, eventually to exceed our own high intelligence and become a super-loud intelligence — a roar! — way beyond us, and maybe even off the chart.

This model is topologically equivalent to a ladder, so that each rung of intelligence is a step higher than the one before. Inferior animals are situated on lower rungs below us, while higher-level intelligence AIs will inevitably overstep us onto higher rungs. Time scales of when it happens are not important; what is important is the ranking—the metric of increasing intelligence.

The problem with this model is that it is mythical, like the ladder of evolution. The pre-Darwinian view of the natural world

supposed a ladder of being, with inferior animals residing on rungs below human. Even post-Darwin, a very common notion is the "ladder" of evolution, with fish evolving into reptiles, then up a step into mammals, up into primates, into humans, each one a little more evolved (and of course smarter) than the one before it. So the ladder of intelligence parallels the ladder of existence. But both of these models supply a thoroughly unscientific view.



A more accurate chart of the natural evolution of species is a disk radiating outward, like this one (above) first devised by David Hillis at the University of Texas and based on DNA. This deep genealogy mandala begins in the middle with the most primeval life forms, and then branches outward in time. Time moves outward so that the most recent species of life living on the planet today form the perimeter of the circumference of this circle. This picture emphasizes a fundamental fact of evolution that is hard to appreciate: Every species alive today is equally evolved. Humans exist on this outer ring alongside cockroaches, clams, ferns, foxes, and bacteria. Every one of these species has undergone an unbroken chain of three billion years of successful reproduction, which means that bacteria and cockroaches today are as highly evolved as humans. There is no ladder.

Likewise, there is no ladder of intelligence. Intelligence is not a single dimension. It is a complex of many types and modes of cognition, each one a continuum. Let's take the very simple task of measuring animal intelligence. If intelligence were a single dimension we should be able to arrange the intelligences of a parrot, a dolphin, a horse, a squirrel, an octopus, a blue whale, a cat, and a gorilla in the correct ascending order in a line. We currently have no scientific evidence of such a

line. One reason might be that there is no difference between animal intelligences, but we don't see that either. Zoology is full of remarkable differences in how animals think. But maybe they all have the same relative "general intelligence?" It could be, but we have no measurement, no single metric for that intelligence. Instead we have many different metrics for many different types of cognition.

Instead of a single decibel line, a more accurate model for intelligence is to chart its possibility space, like the above rendering of possible forms created by an algorithm written by Richard Dawkins. Intelligence is a combinatorial continuum. Multiple nodes, each node a continuum, create complexes of high diversity in high dimensions. Some intelligences may be very complex, with many sub-nodes of thinking. Others may be simpler but more extreme, off in a corner of the space. These complexes we call intelligences might be thought of as symphonies comprising many types of instruments. They vary not only in loudness, but also in pitch, melody, color, tempo, and so on. We could think of them as ecosystem. And in that sense, the different component nodes of thinking are co-dependent, and co-created.

Human minds are societies of minds, in the words of Marvin Minsky. We run on ecosystems of thinking. We contain multiple species of cognition that do many types of thinking: deduction, induction, symbolic reasoning, emotional intelligence, spacial logic, short-term memory, and long-term memory. The entire nervous system in our gut is also a type of brain with its own mode of cognition. We don't really think with just our brain; rather, we think with our whole bodies.

These suites of cognition vary between individuals and between species. A squirrel can remember the exact location of several thousand acorns for years, a feat that blows human minds away. So in that one type of cognition, squirrels exceed humans. That superpower is bundled with some other modes that are dim compared to ours in order to produce a squirrel mind. There are many other specific feats of cognition in the animal kingdom that are superior to humans, again bundled into different systems.

Likewise in AI. Artificial minds already exceed humans in certain dimensions. Your calculator is a genius in math; Google's memory is already beyond our own in a certain dimension. We are engineering AIs to excel in specific modes. Some of these modes are things we can do, but they can do better, such as probability or math. Others are type of thinking we can't do at all — memorize every single word on six billion web pages, a feat any search engine can do. In the future, we will invent whole new modes of cognition that don't exist in us and don't exist anywhere in biology. When we invented artificial flying we were inspired by biological modes of flying, primarily flapping wings. But the flying we invented — propellers bolted to a wide fixed wing — was a new mode of flying unknown in our biological world. It is alien flying. Similarly, we will invent whole new modes of thinking that do not exist in nature. In many cases they will be new, narrow, "small," specific modes for specific jobs — perhaps a type of reasoning only useful in statistics and probability.

In other cases the new mind will be complex types of cognition that we can use to solve problems our intelligence alone cannot. Some of the hardest problems in business and science may require a two-step solution. Step one is: Invent a new mode of thought to work with our minds. Step two: Combine to solve the problem. Because we are solving problems we could not solve before, we want to call this cognition "smarter" than us, but really it is different than us. It's the differences in thinking that are the main benefits of AI. I think a useful model of AI is to think of it as alien intelligence (or artificial aliens). Its alienness will be its chief asset.

At the same time we will integrate these various modes of cognition into more complicated, complex societies of mind. Some of these complexes will be more complex than us, and because they will be able to solve problems we can't, some will want to call them superhuman. But we don't call Google a superhuman AI even though its memory is beyond us, because there are many things we can do better than it. These complexes of artificial intelligences will for sure be able to exceed us in many dimensions, but no one entity will do all we do better. It's similar to the physical powers of humans. The industrial revolution is 200 years old, and while all machines as a class can beat the physical achievements of an individual human (speed of running, weight lifting, precision cutting, etc.), there is no one machine that can beat an average human in everything he or she does.

Even as the society of minds in an AI become more complex, that complexity is hard to measure scientifically at the moment. We don't have good operational metrics of complexity that could determine whether a cucumber is more complex than a Boeing 747, or the ways their complexity might differ. That is one of the reasons why we don't have good metrics for smartness as well. It will become very difficult to ascertain whether mind A is more complex than mind B, and for the same reason to declare whether mind A is smarter than mind B. We will soon arrive at the obvious realization that "smartness" is not a single dimension, and that what we really care about are the many other ways in which intelligence operates — all the other nodes of cognition we have not yet discovered.

2.

The second misconception about human intelligence is our belief that we have a general purpose intelligence. This repeated belief influences a commonly stated goal of AI researchers to create an artificial general purpose intelligence (AGI). However, if we view intelligence as providing a large possibility space, there is no general purpose state. Human intelligence is not in some central position, with other specialized intelligence revolving around it. Rather, human intelligence is a very, very specific type of intelligence that has evolved over many millions of years to enable our species to survive on this planet. Mapped in the space of all possible intelligences, a human-type of intelligence will be stuck in the corner somewhere, just as our world is stuck at the edge of vast galaxy.

We can certainly imagine, and even invent, a Swiss-army knife type of thinking. It kind of does a bunch of things okay, but none of them very well. AIs will follow the same engineering maxim that all things made or born must follow: You cannot optimize every dimension. You can only have tradeoffs. You can't have a general multi-purpose unit outperform specialized functions. A big "do everything" mind can't do everything as well as those things done by specialized agents. Because we believe our human minds are general purpose, we tend to believe that cognition does not follow the engineer's tradeoff, that it will be possible to build an intelligence that maximizes all modes of thinking. But I see no evidence of that. We simply haven't invented enough varieties of minds to see the full space (and so far we have tended to dismiss animal minds as a singular type with variable amplitude on a single dimension.)

3.

Part of this belief in maximum general-purpose thinking comes from the concept of universal computation. Formally described as the Church-Turing hypothesis in 1950, this conjecture states that all computation that meets a certain threshold is equivalent. Therefore there is a universal core to all computation, whether it occurs in one machine with many fast parts, or slow parts, or even if it occurs in a biological brain, it is the same logical process. Which means that you should be able to emulate any computational process (thinking) in any machine that can do "universal" computation. Singularitans rely on this principle for their expectation that we will be able to engineer silicon brains to hold human minds, and that we can make artificial minds that think like humans, only much smarter. We should be skeptical of this hope because it relies on a misunderstanding of the Church-Turing hypothesis.

The starting point of the theory is: "Given infinite tape [memory] and time, all computation is equivalent." The problem is that in reality, no computer has infinite memory or time. When you are operating in the real world, real time makes a huge difference, often a life-or-death difference. Yes, all thinking is equivalent if you ignore time. Yes, you can emulate human-type thinking in any matrix you want, as long as you ignore time or the real-life constraints of storage and memory.

However, if you incorporate time, then you have to restate the principal in a significant way: Two computing systems operating on vastly different platforms won't be equivalent in real time. That can be restated again as: The only way to have equivalent modes of thinking is to run them on equivalent substrates. The physical matter you run your computation on — particularly as it gets more complex — greatly influences the type of cognition that can be done well in real time.

I will extend that further to claim that the only way to get a very human-like thought process is to run the computation on very human-like wet tissue. That also means that very big, complex artificial intelligences run on dry silicon will produce big, complex, unhuman-like minds. If it would be possible to build artificial wet brains using human- like grown neurons, my prediction is that their thought will be more similar to ours. The benefits of such a wet brain are proportional to how similar we make the substrate. The costs of creating wetware is huge and the closer that tissue is to human brain tissue, the more costefficient it is to just make a human. After all, making a human is something we can do in nine months.

Furthermore, as mentioned above, we think with our whole bodies, not just with our minds. We have plenty of data showing how our gut's nervous system guides our "rational" decision-making processes, and can predict and learn. The more we model the entire human body system, the closer we get to replicating it. An intelligence running on a very different body (in dry silicon instead of wet carbon) would think differently.

I don't see that as a bug but rather as a feature. As I argue in point 2, thinking differently from humans is AI's chief asset. This is yet another reason why calling it "smarter than humans" is misleading and misguided.

4.

At the core of the notion of a superhuman intelligence — particularly the view that this intelligence will keep improving itself — is the essential belief that intelligence has an infinite scale. I find no evidence for this. Again, mistaking intelligence as a single dimension helps this belief, but we should understand it as a belief. There is no other physical dimension in the

universe that is infinite, as far as science knows so far.

Temperature is not infinite — there is finite cold and finite heat. There is finite space and time. Finite speed. Perhaps the mathematical number line is infinite, but all other physical attributes are finite. It stands to reason that reason itself is finite, and not infinite. So the question is, where is the limit of intelligence? We tend to believe that the limit is way beyond us, way "above" us, as we are "above" an ant. Setting aside the recurring problem of a single dimension, what evidence do we have that the limit is not us? Why can't we be at the maximum? Or maybe the limits are only a short distance away from us? Why do we believe that intelligence is something that can continue to expand forever?

A much better way to think about this is to see our intelligence as one of a million types of possible intelligences. So while each dimension of cognition and computation has a limit, if there are hundreds of dimensions, then there are uncountable varieties of mind — none of them infinite in any dimension. As we build or encounter these uncountable varieties of mind we might naturally think of some of them as exceeding us. In my recent book The Inevitable, I sketched out some of that variety of minds that were superior to us in some way. Here is an incomplete list:

Some folks today may want to call each of these entities a superhuman AI, but the sheer variety and alienness of these minds will steer us to new vocabularies and insights about intelligence and smartness.

Second, believers of Superhuman AI assume intelligence will increase exponentially (in some unidentified single metric), probably because they also assume it is already expanding exponentially. However, there is zero evidence so far that intelligence — no matter how you measure it — is increasing exponentially. By exponential growth I mean that artificial intelligence doubles in power on some regular interval. Where is that evidence? Nowhere I can find. If there is none now, why do we assume it will happen soon? The only thing expanding on an exponential curve are the inputs in AI, the resources devoted to producing the smartness or intelligences. But the output performance is not on a Moore's law rise. AIs are not getting twice as smart every 3 years, or even every 10 years.

I asked a lot of AI experts for evidence that intelligence performance is on an exponential gain, but all agreed we don't have metrics for intelligence, and besides, it wasn't working that way. When I asked Ray Kurzweil, the exponential wizard himself, where the evidence for exponential AI was, he wrote to me that AI does not increase explosively but rather by levels. He said: "It takes an exponential improvement both in computation and algorithmic complexity to add each additional level to the hierarchy…. So we can expect to add levels linearly because it requires exponentially more complexity to add each additional layer, and we are indeed making exponential progress in our ability to do this. We are not that many levels away from being comparable to what the neocortex can do, so my 2029 date continues to look comfortable to me."

What Ray seems to be saying is that it is not that the power of artificial intelligence is exploding exponentially, but that the effort to produce it is exploding exponentially, while the output is merely raising a level at a time. This is almost the opposite of the assumption that intelligence is exploding. This could change at some time in the future, but artificial intelligence is clearly not increasing exponentially now.

Therefore when we imagine an "intelligence explosion," we should imagine it not as a cascading boom but rather as a scattering exfoliation of new varieties. A Cambrian explosion rather than a nuclear explosion.

The results of accelerating technology will most likely not be super-human, but extra-human. Outside of our experience, but not necessarily "above" it.

5.

Another unchallenged belief of a super AI takeover, with little evidence, is that a super, near-infinite intelligence can quickly solve our major unsolved problems.

Many proponents of an explosion of intelligence expect it will produce an explosion of progress. I call this mythical belief "thinkism." It's the fallacy that future levels of progress are only hindered by a lack of thinking power, or intelligence. (I might also note that the belief that thinking is the magic super ingredient to a cure-all is held by a lot of guys who like to think.)

Let's take curing cancer or prolonging longevity. These are problems that thinking alone cannot solve. No amount of thinkism will discover how the cell ages, or how telomeres fall off. No intelligence, no matter how super duper, can figure out how the human body works simply by reading all the known scientific literature in the world today and then contemplating it. No super AI can simply think about all the current and past nuclear fission experiments and then come up

with working nuclear fusion in a day. A lot more than just thinking is needed to move between not knowing how things work and knowing how they work.

There are tons of experiments in the real world, each of which yields tons and tons of contradictory data, requiring further experiments that will be required to form the correct working hypothesis. Thinking about the potential data will not yield the correct data.

Thinking (intelligence) is only part of science; maybe even a small part. As one example, we don't have enough proper data to come close to solving the death problem. In the case of working with living organisms, most of these experiments take calendar time. The slow metabolism of a cell cannot be sped up. They take years, or months, or at least days, to get results. If we want to know what happens to subatomicparticles, we can't just think about them. We have to build very large, very complex, very tricky physical structures to find out. Even if the smartest physicists were 1,000 times smarter than they are now, without a Collider, they will know nothing new.

There is no doubt that a super AI can accelerate the process of science. We can make computer simulations of atoms or cells and we can keep speeding them up by many factors, but two issues limit the usefulness of simulations in obtaining instant progress. First, simulations and models can only be faster than their subjects because they leave something out. That is the nature of a model or simulation. Also worth noting:

The testing, vetting and proving of those models also has to take place in calendar time to match the rate of their subjects. The testing of ground truth can't be sped up. These simplified versions in a simulation are useful in winnowing down the most promising paths, so they can accelerate progress. But there is no excess in reality; everything real makes a difference to some extent; that is one definition of reality. As models and simulations are beefed up with more and more detail, they come up against the limit that reality runs faster than a 100 percent complete simulation of it. That is another definition of reality: the fastest possible version of all the details and degrees of freedom present. If you were able to model all the molecules in a cell and all the cells in a human body, this simulation would not run as fast as a human body. No matter how much you thought about it, you still need to take time to do experiments, whether in real systems or in simulated systems.

To be useful, artificial intelligences have to be embodied in the world, and that world will often set their pace of innovations. Without conducting experiments, building prototypes, having failures, and engaging in reality, an intelligence can have thoughts but not results. There won't be instant discoveries the minute, hour, day or year a so-called "smarter-than-human" AI appears. Certainly the rate of discovery will be significantly accelerated by AI advances, in part because alien-ish AI will ask questions no human would ask, but even a vastly powerful (compared to us) intelligence doesn't mean instant progress. Problems need far more than just intelligence to be solved.

Not only are cancer and longevity problems that intelligence alone can't solve, so is intelligence itself. The common trope among Singularitans is that once you make an AI "smarter than humans" then all of sudden it thinks hard and invents an AI "smarter than itself," which thinks harder and invents one yet smarter, until it explodes in power, almost becoming god-like. We have no evidence that merely thinking about intelligence is enough to create new levels of intelligence. This kind of thinkism is a belief. We have a lot of evidence that in addition to great quantities of intelligence we need experiments, data, trial and error, weird lines of questioning, and all kinds of things beyond smartness to invent new kinds of successful minds.

I'd conclude by saying that I could be wrong about these claims. We are in the early days. We might discover a universal metric for intelligence; we might discover it is infinite in all directions. Because we know so little about what intelligence is (let alone consciousness), the possibility of some kind of AI singularity is greater than zero. I think all the evidence suggests that such a scenario is highly unlikely, but it is greater than zero.

So while I disagree on its probability, I am in agreement with the wider aims of OpenAI and the smart people who worry about a superhuman AI — that we should engineer friendly AIs and figure out how to instill self-replicating values that match ours. Though I think a superhuman AI is a remote possible existential threat (and worthy of considering), I think its unlikeliness (based on the evidence we have so far) should not be the guide for our science, policies, and development. An asteroid strike on the Earth would be catastrophic. Its probability is greater than zero (and so we should support the B612 Foundation), but we shouldn't let the possibility of an asteroid strike govern our efforts in, say, climate change, or space travel, or even city planning.

Likewise, the evidence so far suggests AIs most likely won't be superhuman but will be many hundreds of extra-human new species of thinking, most different from humans, none that will be general purpose, and none that will be an instant god solving major problems in a flash. Instead there will be a galaxy of finite intelligences, working in unfamiliar dimensions, exceeding our thinking in many of them, working together with us in time to solve existing problems and create new problems.

I understand the beautiful attraction of a superhuman AI god. It's like a new Superman. But like Superman, it is a mythical figure. Somewhere in the universe a Superman might exist, but he is very unlikely. However myths can be useful, and once invented they won't go away. The idea of a Superman will never die. The idea of a superhuman AI Singularity, now that it has been birthed, will never go away either. But we should recognize that it is a religious idea at this moment and not a scientific one. If we inspect the evidence we have so far about intelligence, artificial and natural, we can only conclude that our speculations about a mythical superhuman AI god are just that: myths.

Many isolated islands in Micronesia made their first contact with the outside world during World War II. Alien gods flew over their skies in noisy birds, dropped food and goods on their islands, and never returned. Religious cults sprang up on the islands praying to the gods to return and drop more cargo. Even now, fifty years later, many still wait for the cargo to return. It is possible that superhuman AI could turn out to be another cargo cult. A century from now, people may look back to this time as the moment when believers began to expect a superhuman AI to appear at any moment and deliver them goods of unimaginable value. Decade after decade they wait for the superhuman AI to appear, certain that it must arrive soon with its cargo.

Yet non-superhuman artificial intelligence is already here, for real. We keep redefining it, increasing its difficulty, which imprisons it in the future, but in the wider sense of alien intelligences — of a continuous spectrum of various smartness, intelligences, cognition, reasonings, learning, and consciousness — AI is already pervasive on this planet and will continue to spread, deepen, diversify, and amplify. No invention before will match its power to change our world, and by century's end AI will touch and remake everything in our lives. Still the myth of a superhuman AI, poised to either gift us super-abundance or smite us into super-slavery (or both), will probably remain alive—a possibility too mythical to dismiss.

Art direction by Robert Shaw.

(Newly formatted from the original publication, FN)