Nick Bostrom

Superintelligence: Paths, Dangers, Strategies

Made by Blinkist

These key insights in blinks were written by a team of experts at Blinkist. We screen the world of nonfiction to choose the very best books. Then, we read them deeply and transform them into this concise format that brings you the most inspiring ideas from the text.

Maybe these blinks will inspire you to dig deeper, or maybe they're enough to start you thinking and then on to something new. However you read blinks, we hope they help you become an even brighter you.

What's in it for me? Learn how machines may surpass humanity.

How many movies, cartoons and sci-fi series have you seen featuring some kind of superintelligent robotic race? Probably quite a few. In some films, such as *Terminator,* they come to conquer the world; in others, they help us out; and in some, like *Wall-E*, they're simply adorable. Of course, these robots are fictional, but will they always be? Will the future bring superintelligent AI? If it does, what will they look like and when will they appear?

In *Superintelligence* we learn about the journey toward AI so far – where we might be going; the moral issues and safety concerns we need to address; and the best ways to reach the goal of creating a machine that'll outsmart all others.

In these blinks, you'll discover

- why most scientists believe we will reach superintelligence by 2105;
- the difference between Artificial Intelligence and Whole Brain Emulation; and
- how a 1956 conference in Dartmouth played a central role in creating the technology.

History shows that superintelligence – a technology more intelligent than any human being – is fast approaching.

What fundamentally sets us apart from the beasts of the field? Well, the main difference between human beings and animals is our capacity for abstract thinking paired with the ability to communicate and accumulate information. In essence, our superior intelligence propelled us to the top.

So what would the emergence of a new species, intellectually superior to humans, mean for the world?

First we'll need to review a bit of history. For instance, did you know that the pace of major revolutions in technology has been increasing over time? For example, improving at the snail's pace of a few hundred thousand years ago, human technology would have needed one million years to become economically

productive enough to sustain the lives of an additional million people. This number dropped to two centuries during the Agricultural Revolution in 5,000 BC. And in our post-Industrial Revolution era it shrunk to a mere 90 minutes.

A technological advancement like the advent of superintelligent machines would mean radical change for the world as we know it. But where does technology stand at present?

We have already been able to create machines that have the capacity to learn and reason using information that's been plugged in by humans. Consider, for example, the automated spam filters that keep our inboxes free from annoying mass emails and save important messages.

However, this is far from the kind of "general intelligence" humans possess, and which has been the goal of AI research for decades. And when it comes

to building a superintelligent machine that can learn and act without the guiding hand of a human, it may still be decades away. But advancements in the field are happening quickly, so it could be upon us faster than we think. Such a machine would have a lot of power over our lives. Its intelligence could even be dangerous, since it would be too smart for us to disable in the event of an emergency.

The history of machine intelligence over the past half decade has had its ups and downs.

Since the invention of computers in 1940, scientists have been working to build a machine that can think. What progress has been made? One major advance in Artificial Intelligence (or AI) are man-made machines that mimic our own intelligence.

The story begins with the 1956 Dartmouth Summer Project, which endeavored to build intelligent machines that could do what humans do.

Some machines could solve calculus problems, while others could write music and even drive cars. But there was a roadblock: inventors realized that the more complex the task, the more information the AI needed to process. Hardware to take on such difficult functions was unavailable.

By the mid-1970s, interest in AI had faded. But in the early '80s, Japan developed *expert systems* – rule-based programs that helped decision-makers by generating inferences based on data. However, this technology also encountered a problem: the huge banks of information required proved difficult to maintain, and interest dropped once again.

The '90s witnessed a new trend: machines that mimicked human biology by using technology to copy our neural and genetic structures. This process brings us up to the present day. Today, AI is present in everything from robots that conduct surgeries to smartphones to a simple Google search. The technology has improved to the point where it can beat the best human players at chess, Scrabble and *Jeopardy!*

But even our modern technology has issues: such AIs can only be programmed

for one game and there's no AI capable of mastering *any* game.

However, our children may see something much more advanced – the advent of *superintelligence* (or SI). In fact, according to a survey of international experts at The Second Conference on Artificial General Intelligence at the University of Memphis, in 2009, most experts think that machines as intelligent as humans will exist by 2075 and that superintelligence will exist within another 30 years.

Superintelligence is likely to emerge in two different ways.

It's clear that imitating human intelligence is an effective way to build technology, but imitation comes in many forms. So, while some scientists are in favor of synthetically designing a machine that simulates humans (through AI, for instance), others stand by an exact imitation of human biology, a strategy that could be accomplished with techniques like *Whole Brain Emulation* (or WBE).

So what are the differences between the two?

AI mimics the way humans learn and think by calculating probability. Basically, AI uses logic to find simpler ways of imitating the complex abilities of humans. For instance, an AI programmed to play chess chooses the optimal move by first determining all possible moves

and then picking the one with the highest probability of winning the game. But this strategy relies on a data bank that holds every possible chess move.

Therefore, an AI that does more than just play chess would need to access and process huge amounts of real world information. The problem is that present computers just can't process the necessary amount of data fast enough.

But are there ways around this?

One potential solution is to build what the computer scientist Alan Turing called "the child machine," a computer that comes with basic information and is designed to learn from experience.

Another option is WBE, which works by replicating the entire neural structure of the human brain to imitate its function. One advantage this method has over AI is that it doesn't require a complete understanding of the processes behind

the human brain – only the ability to duplicate its parts and the connections between them.

For instance, a scientist could take a stabilized brain from a corpse, fully scan it, then translate that information into code. But we'll have to wait. The technology necessary for this process – high-precision brain scans, for instance – likely won't be developed anytime soon. But, someday, it *will*.

“The fact that there are many paths that lead to superintelligence should increase our confidence that we will eventually get there.”

Superintelligence will either emerge quickly via strategic dominance or as a result of long collaborative efforts.

Most of the great discoveries of humanity were achieved either by a single scientist who reached a goal before others got there or through huge international collaborations. So, what would each route mean for the development of SI?

Well, if a single group of scientists were to rapidly find solutions to the issues preventing AI and WBE, it's most likely their results would produce a *single* superintelligent machine. That's because the field's competitive nature might force such a group to work in secrecy.

Consider the Manhattan Project, the group that developed the atom bomb. The group's activities were kept secret because the U.S. government feared that

the USSR would use their research to build nuclear weapons of their own.

If SI developed like this, the first superintelligent machine would have a strategic advantage over all others. The danger is that a single SI might fall into nefarious hands and be used as a weapon of mass destruction. Or if a machine malfunctioned and tried to do something terrible – kill all humans, say – we'd have neither the intelligence nor the tools necessary to defend ourselves.

However, if multiple groups of scientists collaborated, sharing advances in technology, humankind would *gradually* build SI. A team effort like this might involve many scientists checking every step of the process, ensuring that the best choices have been made.

A good precedent for such collaboration is the Human Genome Project, an effort that brought together scientists from multiple countries to map human DNA.

Another good technique would be public oversight – instating government safety regulations and funding stipulations that deter scientists from working independently.

So, while the rapid development of a single SI could still occur during such a slow collaborative process, an open team effort would be more likely to have safety protocols in place.

We can prevent unintended catastrophes by programing superintelligence to learn human values.

You've probably heard it a million times, but there is some wisdom in being careful what you wish for. While we may be striving to attain superintelligence, how can we ensure that the technology doesn't misunderstand its purpose and cause unspeakable devastation?

The key to this problem lies in programing the motivation for SI to accomplish its various human-given goals. Say we designed an SI to make paper clips; it seems benign, but what's to prevent the machine from taking its task to an extreme and sucking up all the world's resources to manufacture a mountain of office supplies?

This is tricky, because while AI is only motivated to achieve the goal for which it has been programmed, an SI would

likely go beyond its programmed objectives in ways that our inferior minds couldn't predict.

But there are solutions to this problem. For instance, superintelligence, whether it be AI or WBE, can be programmed to learn the values of a human on its own. For example, an SI could be taught to determine whether an action is in line with a core human value. In this way we could program SI to do things like "minimize unnecessary suffering" or "maximize returns."

Then, before acting, the machine would calculate whether a proposed action is in line with that goal. With experience, the AI would develop a sense of which actions are in compliance and which aren't.

But there's another option. We could also program an AI to infer our intentions based on the majority values of human beings.

Here's how:

The AI would watch human behavior and determine normative standards for human desires. The machine would essentially be programmed to program itself. For instance, while each culture has it's own culinary tradition, all people agree that poisonous foods should not be eaten. By constantly learning through observation, the SI could self-correct by changing its standards to correspond to changes in the world over time.

Intelligent machines will probably replace the entire human workforce.

But enough about decimation and total destruction. Before panicking about the impending machine-led apocalypse, let's take a look at how general intelligence technology can be developed and put to productive use.

It's likely that the increasing availability and decreasing cost of technology will lead to the cheap mass production of machines capable of doing jobs that currently require the hands and mind of a human. This means that machines will not only replace the entire human workforce but will also be easily replaceable.

For instance, if a WBE worker needs a break, just like a real human would, it can simply be replaced with a fresh unit and no productive time needs to be sacrificed. In fact, it would be easy to do

this by programing a template WBE that thinks it just got back from vacation. This template could then be used to make infinite copies of new workers.

But clearly this amounts to mechanical slavery and raises important moral issues. For example, if a machine became aware that it would die at the end of the day, we could simply program it to embrace death. But is that ethical? Should these artificial employees be treated like sentient beings or inert tools?

Work isn't the only thing that SI machines could take over; they could also be in charge of various mundane tasks in our personal lives. As the minds of these machines come increasingly closer to resembling those of human beings, we could use them to optimize our lives; for instance, we could design a digital program that verbally articulates our thoughts or that achieves our

personal goals better than we could alone.

The result of such advances would mean a human existence that is largely automated, low-risk, devoid of adventure and, frankly, too perfect. And where would that leave us? How would we occupy ourselves in such a future?

“It would be a society of economic miracles and technological awesomeness, with nobody there to benefit. A Disneyland without children.”

In the superintelligent future, the average human will be impoverished or reliant on investments; the rich will be buying new luxuries.

It's clear that an entirely robotic workforce would completely transform the economy, as well as our lifestyles and desires; as machine labor becomes the new, cheaper norm, the pay of workers will drop so low that no human will be able to live off a paycheck. Also, the few employers of the mechanical workforce would accrue *a lot* of money.

But this brings us back to an earlier point, because where that money ends up also depends on whether SI is designed by a single exclusive group or is the result of a slow collaborative process. If the former turns out to be true, most people would be left with few options for income generation, likely renting housing to other humans or relying on their life-savings and pensions.

And the people who don’t have property or savings?

They would be destitute. Their only options would be to use their remaining money to upload themselves into a digital life form, if such technology exists, or rely on charity from the hyper-wealthy.

And the rich?

They’ll lose interest in what we today consider highly desirable luxuries. That’s because with machines doing all the work, anything made or offered by a human will become a highly-valued rarity, much like artisanal products are in our time. While today it might be wine or cheese, in the future it could be something as simple as a handmade key chain.

But the new mode of production would also make possible an unimaginable variety of technological products –

maybe even the ability to live forever or regain youth. So instead of buying yachts and private islands, the wealthy might use their money to upload themselves into digital brains or virtually indestructible humanoid bodies.

However, this scenario assumes that the superintelligent worker robots will not rebel and try to destroy human society. Therefore, whatever route we follow with SI, safety will always be key.

Safety must be a top priority before superintelligence is developed.

It's clear that the development of SI comes with a variety of safety issues and, in the worst case scenario, could lead to the destruction of humankind. While we can take *some* precautions by considering the motivation for the SI we build, that alone won't suffice.

So what will?

Considering every potential scenario *before* bringing a hyper-powerful force like SI into the world. For instance, imagine that some sparrows adopted a baby owl. Having a loyal owl around might be highly advantageous; the more powerful bird could guard the young, search for food and do any number of other tasks. But these great benefits also come with a great risk: the owl might realize it's an owl and eat all the sparrows.

Therefore, the logical approach would be for the sparrows to design an excellent plan for how to teach the owl to love sparrows, while also considering all possible outcomes wherein the owl could become a negative force.

So how can we teach our robotic, superintelligent baby owl to love humans?

As we already know, we can make safety a priority through a long-term international collaboration. But why would the competitive rush to design the first SI be a safety threat?

Because scientists would forgo safety to speed up their process and wouldn't share their work with others. That means that if an SI project went horribly wrong and threatened humanity with extinction, too few people would understand the machine's design well enough to stop it.

On the other hand, if governments, institutions and research groups join together, they could slowly build a safe and highly beneficial SI. That's because groups could share their ideas for safety measures and provide thorough oversight for each phase of the design. Not just that, but an international superintelligence project would promote peace through its universal benefits. Just consider the International Space Station, an endeavor that helped stabilize relations between the US and the USSR.

“Before the prospect of an intelligence explosion, we humans are like small children playing with a bomb.”

Final summary

The key message in this book:

Inventing a superintelligent machine capable of things far beyond the ability of a human is both a tantalizing prospect and a precarious path. In order to ensure such technology develops in a safe, responsible manner, we need to prioritize safety over unchecked technological advancement. The fate of our species depends on it.

Suggested further reading: *Out of Control* by Kevin Kelly

Though written from the perspective of 1994, these blinks paint a startlingly current and still futuristic image of how technological developments like the internet and artificial intelligence could affect society and humanity.

Got feedback?

We’d sure love to hear what you think about our content! Just drop an email to remember@blinkist.com with the title of this book as the subject line and share your thoughts!

Nice work! You're all done with this one.

We publish new books every week at blinkist.com.
Come and see – there's so much more to learn.

Inspired to read the full book?

Get it here.