

Books About AI & Humans (Nonfiction)



Published on: bradleymurray.ca

Contents

Contents.....	2
Life 3.0: Being Human in the Age of Artificial Intelligence, by Max Tegmark.....	4
What is Life?.....	4
Intelligence.....	5
AGI Aftermath.....	6
Libertarian utopia.....	6
Benevolent dictator.....	6
Gatekeeper.....	7
1984.....	7
Conclusion.....	7
Superintelligence: Paths, Dangers, Strategies, by Nick Bostrom.....	7
Definition of Superintelligence.....	8
Anthropomorphism and Superintelligence.....	9
Superintelligence Goals and Motivation.....	10
Problems of Value.....	12
The Common Good Principle.....	13
The Singularity Is Near: When Humans Transcend Biology, by Ray Kurzweil.....	13
Kurzweil's "Singularity" Vision.....	14
Acceleration of Technological Innovation and Paradigm Shifts.....	14
Emulation of the Human Brain and the Development of Nonbiological Intelligence.....	15
Human Enhancement.....	16
Virtual Reality.....	16
Spread of Intelligence.....	17
Human Values.....	17
Consciousness.....	18
Risks and Relinquishment.....	19
Conclusions.....	19

Life 3.0: Being Human in the Age of Artificial Intelligence,

by Max Tegmark

Life 3.0: Being Human in the Age of Artificial Intelligence is an accessible book about artificial intelligence that offers a review of some of the most important ideas being debated among AI researchers as they conceive and implement this novel technology.

What is Life?

The question of how to define "life" has been a source of debate among scientists and philosophers for a very long time. Tegmark does not wade into these debates, but instead makes the decision from the outset to define "life," in its broadest sense, as "a process that can retain its complexity and replicate." He adds to the explanation that what is replicated "isn't matter (made of atoms) but information (made of bits) specifying how the atoms are arranged." Concerning this "information-processing" definition of life, he writes:

[W]e can think of life as a self-replicating information-processing system whose information (software) determines both its behavior and the blueprints for its hardware.

With this definition in mind, Life 3.0 stipulates that life admits of three levels of sophistication.

- Life 1.0: Can survive and replicate; design for its software and hardware are determined by its DNA, and change only through evolution over many generations (simple biological)
- Life 2.0: Can design its software but evolves its hardware; humans can learn complex new skills, including languages, sports and professions, and can fundamentally update

their worldview and

goals (cultural)

- Life 3.0: Can design its hardware; does not yet exist on Earth, can dramatically redesign not only its software, but its hardware as well, rather than having to wait for it to gradually evolve over generations (technological)

Intelligence

Tegmark defines "intelligence" in a very broad way, just as he defines "life" very broadly.

Intelligence, he claims, amounts to the "ability to accomplish complex goals." Accordingly, there can be many possible types of intelligence.

When it comes to artificial general intelligence, or AGI, Tegmark accepts the view that this would be human-level intelligence, or the "ability to accomplish any goal at least as well as humans."

Chapter 3 of Life 3.0 reviews different ways in which Tegmark believes that artificial intelligence is being, or could be, used by humanity. These include:

- Space exploration
- Finance
- Manufacturing
- Transportation
- Energy
- Healthcare
- Secure communication
- Legal system
- Weapons

AGI Aftermath

Among the most engaging and creative sections of Life 3.0 are those which consider in detail scenarios that might emerge for humanity on the path to, or in the aftermath of the development of artificial general intelligence (should this ever occur).

Libertarian utopia

Tegmark describes this as a situation in which humans, cyborgs, uploads and superintelligences coexist peacefully thanks to property rights. It is the scenario that he finds articulated by futurists like Hans Moravec and Ray Kurzweil, in which many humans will have upgraded their bodies or even uploaded their minds into new hardware. The majority of interactions would occur in virtual environments, "for convenience and speed," even if many minds continue to enjoy interactions and activities involving physical bodies as well. Tegmark imagines different "zones," including human-only zones, machine zones, and mixed zones. In the final analysis, Tegmark considers this scenario unlikely, partly because he does not consider it obvious that cyborgs or uploads would ever be made.

Benevolent dictator

In this scenario, a "benevolent superintelligence" is established which "runs the world and enforces strict rules designed to maximize its model of human happiness." Humanity would be free from poverty, disease, and all necessary goods and services would be taken care of. Crime would be eliminated, since the superintelligence would be omniscient and efficient in enforcing social rules. It would be left to the superintelligence to figure out what a human utopia would amount to (and wouldn't bring it about, say, through mass chemical intoxication). Tegmark considers one downside of this scenario to the fact that some people might feel a lack of

freedom in shaping their society and destiny, given the single overall path to human happiness that had been followed in order to maximize human happiness in the aggregate.

Gatekeeper

In this scenario, a superintelligent AI is created with the goal of interfering as little as necessary to prevent the creation of another superintelligence. However, Tegmark imagines critics of such a scenario opposing the obstruction of future technological progress that would rely on further superintelligence.

1984

The "1984" scenario that Tegmark imagines is one in which, for reasons such as to manage technological risks, progress in artificial intelligence comes to a halt because of human intervention. According to Tegmark, this would require an Orwellian surveillance state in which certain forms of AI research are prohibited.

Conclusion

In addition to these scenarios, Tegmark considers several other scenarios, each with its potential benefits and drawbacks. One of the virtues of this book about artificial intelligence is that, without becoming too technical, it helps readers to envision different paths human history might take in light of artificial intelligence technology - and asks readers to begin to think about which is most aligned with their own preferences and values.

Superintelligence: Paths, Dangers, Strategies, by Nick Bostrom

Nick Bostrom's *Superintelligence: Paths, Dangers, Strategies* is a book about AI that examines some of the fundamental challenges that confront us as AI systems continue to be developed and become increasingly intelligent.

This article will not review the specific paths to development and risk that Bostrom analyzes (those interested may consult the book). Instead, the aim here is to discuss some of the book's main themes.

Definition of Superintelligence

Bostrom's book focuses on the possible future emergence of superintelligence. This could occur through different means - but one of the more likely paths is through AI systems.

Bostrom understands machine general intelligence to involve machines matching humans in intelligence, possessing "common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains."

Superintelligence, whether or not it emerges from intelligent machines, is understood as involving intellects that not only match but "greatly outperform the best current human minds across many very general cognitive domains."

Bostrom considers three bundles of intellectual super-capabilities that he counts as superintelligence.

- Speed superintelligence: A system that can do all that a human intellect can do, but much faster. This means being able to do these things multiple orders of magnitude faster, such as reading a book in a few seconds or writing a PhD thesis in an afternoon.

- Collective superintelligence: A system composed of a large number of smaller intellects such that the system's overall performance across many very general domains vastly outstrips that of any current cognitive system. Ordinary examples of collective human-level intelligence already exist, such as firms, work teams, gossip networks, advocacy groups, academic communities, and countries.
- Quality superintelligence: A system that is at least as fast as a human mind and vastly qualitatively smarter. For example, nonhuman animals lack complex structured language, are capable of no or only rudimentary tool use and tool construction, are severely restricted in their ability to make long-term plans, and have very limited abstract reasoning ability. These are differences in intelligence quality from human beings.

Anthropomorphism and Superintelligence

The possibility of the emergence of superintelligence seems not, as Bostrom points out, to have seriously occurred to AI pioneers:

It may seem obvious now that major existential risks would be associated with such an intelligence explosion, and that the prospect should therefore be examined with the utmost seriousness even if it were known (which it is not) to have but a moderately small probability of coming to pass. The pioneers of artificial intelligence, however, notwithstanding their belief in the imminence of human-level AI, mostly did not contemplate the possibility of greater-than-human AI. It is as though their speculation muscle had so exhausted itself in conceiving the radical possibility of machines reaching human intelligence that it could not grasp the corollary—that machines would subsequently become superintelligent.

As Bostrom emphasizes, it would be a mistake to assume that artificial intelligence in general resembles the human mind. They have different cognitive architectures, different profiles of cognitive strengths and weaknesses, and different goal systems.

Just as it is a mistake to anthropomorphize artificial intelligence in general, Bostrom argues, it is a mistake to anthropomorphize potential AI superintelligence. This encourages expectations about the growth trajectory and capabilities of mature superintelligence. According to Bostrom, the tendency towards anthropomorphizing can lead to underestimating the extent to which a machine intelligence could exceed the human level of performance.

Superintelligence Goals and Motivation

Bostrom introduces what he calls the orthogonality thesis:

Intelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal.

For example, just because someone or something is very intelligent, it does not follow that it will act in ways that we would approve of.

Beyond this, even if human programmers specify the goals at the outset, it could be difficult to understand or predict the paths that superintelligence might take in reaching those goals. This is because of the ways in which the goal systems differ from the ways in which humans understand goals. Consequently, there can arise what Bostrom calls "perverse instantiations" in which the system finds a way of satisfying the criteria of the final goal that violates the intentions of the programmers who defined the goal. This becomes an issue for human beings if the superintelligence has by then obtained what Bostrom calls a "decisive strategic advantage."

This is part of a general and very wide-ranging problem that is currently left to programmers. For instance, suppose that a programmer believes that they have discovered a way of specifying a final goal that is not susceptible to a perverse instantiation:

Perhaps it is not immediately obvious how it could have a perverse instantiation. But we should not be too quick to clap our hands and declare victory. Rather, we should worry that the goal specification does have some perverse instantiation and that we need to think harder in order to find it. Even if after thinking as hard as we can we fail to discover any way of perversely instantiating the proposed goal, we should remain concerned that maybe a superintelligence will find a way where none is apparent to us. It is, after all, far shrewder than we are.

Given the nature of the technology, Bostrom suggests, intellectual humility seems the most appropriate attitude:

The claim is that it is much easier to convince oneself that one has found a solution than it is to actually find a solution. This should make us extremely wary. We may propose a specification of a final goal that seems sensible and that avoids the problems that have been pointed out so far, yet which upon further consideration—by human or super-human intelligence—turns out to lead to either perverse instantiation....

This creates a problem of control, which has to do with how to constrain, from the outset, the behavioral paths that might be taken by a future superintelligence in pursuing goals that are set for it. As Bostrom points out, it is essential that this problem be solved before, and not after, the possible emergence of superintelligence with a decisive strategic advantage because (it is presumed) there is a single opportunity to get the goal specification right.

Problems of Value

The next point that *Superintelligence: Paths, Dangers, Strategies* makes is equally important: even if the problem of control does get solved, there remains an important human question of which values are incorporated into the final goal. He writes:

Suppose that we had solved the control problem so that we were able to load any value we chose into the motivation system of a superintelligence, making it pursue that value as its final goal. Which value should we install? The choice is no light matter. If the superintelligence obtains a decisive strategic advantage, the value would determine the disposition of the cosmic endowment. Clearly, it is essential that we not make a mistake in our value selection. But how could we realistically hope to achieve errorlessness in a matter like this? We might be wrong about morality; wrong also about what is good for us; wrong even about what we truly want. Specifying a final goal, it seems, requires making one's way through a thicket of thorny philosophical problems. If we try a direct approach, we are likely to make a hash of things. The risk of mistaken choosing is especially high when the decision context is unfamiliar—and selecting the final goal for a machine superintelligence that will shape all of humanity's future is an extremely unfamiliar decision context if any is.

This, again, suggests the need for humility because these are deep philosophical questions about value that have perplexed leading thinkers for thousands of years. By default, the selection of values would be placed in the hands of the generation of programmers who would build what might evolve into highly advanced intelligence.

Fools might eagerly accept this challenge of solving in one swing all the important problems in moral philosophy, in order to infix their favorite answers into the seed AI. Wiser souls would look hard for some alternative approach, some way to hedge.

The Common Good Principle

Bostrom concludes this book about artificial intelligence by suggesting what he calls the "common good principle," which is that "superintelligence should be developed only for the benefit of all of humanity and in the service of widely shared ethical ideals." While this sounds very good and is certainly worth aspiring to, it is not difficult to see why different people pursuing the principle might take it to amount to very different things. People will disagree over what counts as a "benefit" to humanity; and although there some broadly ethical ideals that may be widely approved of, these can be subject to diverse interpretations in particular cases, leading to widely diverging conceptions of what is good and right.

The Singularity Is Near: When Humans Transcend Biology, by Ray Kurzweil

For those looking to understand the ways in which human life might be impacted by developments in AI and nanotechnology, *The Singularity is Near: When Humans Transcend Biology* is a book about artificial intelligence that provides a thought-provoking vision. While some of its key claims are controversial and uncomfortable to contemplate, it would be a mistake to ignore them. The purpose of this article is to review its claims for readers new to these ideas.

Kurzweil's "Singularity" Vision

Kurzweil defends the view that there will come a point in history he calls the "Singularity" in which human beings become indistinguishable from machines.

The Singularity will represent the culmination of the merger of our biological thinking and existence with our technology, resulting in a world that is still human but that transcends our biological roots. There will be no distinction, post-Singularity, between human and machine or between physical and virtual reality.

He believes that this will be a good thing for humanity, and urges the development of the technologies that he expects will make this possible. This includes a range of technologies which he acknowledges to pose serious risks to humanity, including strong AI and nanotechnology.

According to Kurzweil, the Singularity involves a range of principles, which are summarized below.

Acceleration of Technological Innovation and Paradigm Shifts

One idea that Kurzweil emphasizes early on in *The Singularity is Near* is that technology accelerates faster than many of us - including experts - predict. This is because the power, price-performance, speed, capacity, and bandwidth of information technologies grows exponentially. This accelerating pace of technological development means that the rate of "paradigm shifts," which occur when we fundamentally rethink our understanding of life, accelerates as well. While the Singularity may seem like an event in the distant future, Kurzweil argues that it will creep up on us faster than we would expect. In his book *The Singularity is*

Near: When Humans Transcend Biology, Kurzweil predicts that the date for the Singularity will occur in the 2040s:

I set the date for the Singularity - representing a profound and disruptive transformation in human capability - as 2045. The nonbiological intelligence created in that year will be one billion times more powerful than all human intelligence today.

Emulation of the Human Brain and the Development of Nonbiological Intelligence

Kurzweil takes a strong interest in human brain scanning as a technology that is laying the foundation for the fundamental shifts he expects to occur in the coming decades. The project he believes is needed is to reverse engineer the human brain so that advanced hardware of the future will be able to emulate human intelligence. He predicts that computers will pass the Turing test by the end of the 2020s, which is the standard of machine intelligence that is indistinguishable from that of biological humans. Kurzweil anticipates that computers will be able to combine the strengths of human and machine intelligence, to be able to perform a broad range of functions, including:

- recognizing patterns based on subtle, invariant properties
- learning new knowledge by applying insights and inferring principles from experience
- creating mental models of reality and conducting 'what if' experiments by varying aspects of the models
- remembering billions of facts precisely and recalling them instantly
- performing repetitive skills repeatedly with accuracy and at high speed without tiring
- sharing knowledge at an extremely high speed
- downloading skills from other machines and from humans

- having access via the internet to all the knowledge of human-machine civilization

Human Enhancement

Another area that Kurzweil argues will be crucial in the coming decades is human enhancement. He expects that we will be able to reengineer all the organs and systems in the human body and brain to make them more capable and resilient. He argues that these developments will come about using nanotechnology, which will make possible the manipulation of physical reality at the molecular level. He writes:

Nanotechnology will enable the design of nanobots: robots designed at the molecular level, measured in microns (millionths of a meter), such as "respirocytes" (mechanical red-blood cells). Nanobots will have myriad roles within the human body, including reversing human aging (to the extent that this task will not already have been completed through biotechnology, such as genetic engineering).

He sees what many will find to be a far-fetched human enhancement possibilities becoming reality:

- the use of nanobots which interact with biological neurons to extend human experience by creating virtual reality from within the nervous systems
- billions of nanobots in the capillaries of the brain that extend human intelligence
- the machine intelligence in our brains growing exponentially and at least doubling in power each year

Virtual Reality

The Singularity is Near: When Humans Transcend Biology argues that, as human enhancement unfolds, we can expect more of human experience to occur in virtual reality. Virtual reality environments accessed from within the nervous system will, Kurzweil argues, become competitive with real reality in terms of resolution and believability. As our experiences increasingly take place in virtual environments, our interaction with machine intelligence with virtual bodies will also occur in these environments.

Spread of Intelligence

One of the more far-reaching - and arguably "quasi-religious" - aspects of Kurzweil's program is that he upholds the value of intelligence "spreading" throughout the universe to "saturate" it:

Ultimately, the entire universe will become saturated with our intelligence. This is the destiny of the universe. We will determine our own fate rather than have it determined by the current "dumb," simple, machinelike forces that rule celestial mechanics."

This is a telling passage, as it suggests that there is a kind of enthusiasm that drives Kurzweil's investigations that goes beyond a scientific interest. It illustrates the ways in which the shape of the technological advancement that he foresees (and that has been occurring since the book was released in 2005) comes down to matters of value, and of the unique personalities of those who are driving these technological changes.

Human Values

We further see the connection between individual values and technological change as we read the more philosophical segments of *The Singularity is Near: When Humans Transcend Biology*. Kurzweil presents (philosophically controversial) views about the value of human life and about

death. He regards a person as a "as a profound pattern (a form of knowledge)." He claims that the purpose of life in general, and our lives in particular, is to "create and appreciate ever-greater knowledge, to move toward greater 'order.'" The order that he has in mind typically takes the form of "increasing complexity," but can also take the form of a reduction of complexity as a result of a "profound insight." When it comes to death, Kurzweil claims that the "pattern" that is the person is lost, since we do not yet have the means to "back up this knowledge." He expects that future technologies will enable the uploading and back up of such knowledge, thereby transforming our understanding of death.

Consciousness

Another area of philosophical controversy that Kurzweil enters is that of consciousness. He asks whether future machines will be capable of conscious experience. He writes:

They—we—will claim to be human and to have the full range of emotional and spiritual experiences that humans claim to have. And these will not be idle claims; they will evidence the sort of rich, complex, and subtle behavior associated with such feelings. But how will these claims and behaviors—compelling as they will be— relate to the subjective experience of nonbiological humans?

Kurzweil argues that there is no objective test to determine conclusively whether consciousness is present. Answering such a question is beyond the objective means of science, he claims, since "the very nature of objectivity is that you cannot measure subjective." We are limited to being able to measure correlates of consciousness, such as behavior. Even though we will not be able to definitively answer the question of machine consciousness, Kurzweil suggests that the question will remain essential. This is because the legal and moral foundations of society are built on the idea of human beings' having subjective feelings that need to be respected.

Risks and Relinquishment

The Singularity is Near: When Humans Transcend Biology acknowledges risks of the technologies of the future. For example, nanotechnology carries with it existential risks to humanity, coming from the possibility of hostile actors and self-replicating nanobots that become widely destructive. Kurzweil argues that full relinquishment of the pursuit of these technologies is not desirable. This would be incompatible with his valuing of the Singularity state of affairs he would like to bring about. Rather, he argues for "fine-grained" relinquishment. For example, there is the suggestion that nanotechnologists should agree to "relinquish the development of physical entities that can self-replicate in a natural environment." However, Kurzweil is hesitant to go even this far, since he argues that to guard against the existential risks there will need to be a "nanotechnology-based planetary immune system" in the natural environment, which would itself need to be self-replicating in order to be able to perform the intended function.

Conclusions

Most of us have been struck by the power of existing artificial intelligence technology. Ray Kurzweil's *The Singularity is Near: When Humans Transcend Biology* is a sweeping and forward-thinking look at the way the future might unfold if this technology comes into closer contact with human beings and the line between humans and machines blurs in new and unexpected ways.