En mémoire de Phil Koenig
(25 mai 1964 – 11 juillet 2018)

1.- De quoi exercer les activités de vos cellules grises pendant l'été en pensant à Phil cf below – à la fin

2. Pour en savoir plus

Life 3.0 - Being human in the Age of Artificial Intelligence by Tegmark Max

3.- Pour en savoir beaucoup plus Achetez sans délai le bouquin

4. **Lisez-le** ⬅==========

encore plus

https://www.youtube.com/watch?v=Gi8LUnhP5yU

et sur Wikipedia *Life 3.0*

From Wikipedia, the free encyclopedia
Jump to navigation Jump to search

Life 3.0: Being Human in the Age of Artificial Intelligence

Hardcover edition

Author	Max Tegmark
Country	United States
Language	English
Subject	Artificial Intelligence
Genre	Non-fiction
Publisher	Knopf
Publication date	August 23, 2017
Mediatype	Print (hardback)
Pages	280
ISBN	978-1-101-94659-6

Life 3.0: Being Human in the Age of Artificial Intelligence[1] is a book by Swedish-American cosmologist Max Tegmark from MIT. *Life 3.0* discusses Artificial Intelligence (AI) and its impact on the future of life on Earth and beyond. The book discusses a variety of societal implications, what can be done to maximize the chances of a positive outcome, and potential futures for humanity, technology and combinations thereof.

Contents

Summary

The book begins by positing a scenario in which AI has exceeded human intelligence and become pervasive in society. Tegmark refers to different stages of human life since its inception: Life 1.0 referring to biological origins, Life 2.0 referring to cultural developments in humanity, and Life 3.0 referring to the technological age of humans. The book focuses on "Life 3.0", and on emerging technology such as artificial general intelligence that may someday, in addition to being able to learn, be able to also redesign its own hardware and internal structure.

The first part of the book looks at the origin of intelligence billions of years ago and goes on to project the future development of intelligence. Tegmark considers short-term effects of the development of advanced technology, such as technological unemployment, AI weapons, and the quest for human-level AGI (Artificial General Intelligence). The book cites examples like Deepmind and OpenAI, self-driving cars, and AI players that can defeat humans in Chess,[2] Jeopardy,[3] and Go.[4]

After reviewing current issues in AI, Tegmark then considers a range of possible futures that feature intelligent machines and/or humans. The fifth chapter describes a number of potential outcomes that could occur, such altered social structures, integration of humans and machines, and both positive and negative scenarios like Friendly AI or an AI apocalypse.[5] Tegmark argues that the risks of AI come not from malevolence or conscious behavior per se, but rather from the misalignment of the goals of AI with those of humans. Many of the goals of the book align with those of the Future of Life Institute.[6]

The remaining chapters explore concepts in physics, goals, consciousness and meaning, and investigate what society can do to help create a desirable future for humanity.

Marketing

Business magnate Elon Musk, who had previously endorsed the thesis that, under some scenarios, advanced AI could jeopardize human survival, recommended Life 3.0 as "worth reading".[7][8]

Reception

Professor Max Tegmark, author of *Our Mathematical Universe*.

One criticism of the book by *Kirkus Reviews* is that some of the scenarios or solutions in the book are a stretch or somewhat prophetic: "Tegmark's solutions to inevitable mass unemployment are a stretch."[9] AI researcher Stuart J. Russell, writing in Nature, said: "I am unlikely to disagree strongly with the premise of *Life 3.0*. Life, Tegmark argues, may or may not spread through the Universe and 'flourish for billions or trillions of years' because of decisions we make now — a possibility both seductive and overwhelming."[10] Writing in *Science*, Haym Hirsh called it "a highly readable book that complements The Second Machine Age's economic perspective on the near-term implications of recent accomplishments in AI and the more detailed analysis of how we might get from where we are today to AGI and even the superhuman AI in Superintelligence."[11] *The Telegraph* called it "One of the very best overviews of the arguments around artificial intelligence".[12][13] The Christian Science Monitor said "Although it's probably not his intention, much of what Tegmark writes will quietly terrify his readers."[14] *Publishers Weekly* gave a positive review, but also stated that Tegmark's call for researching how to maintain control over superintelligent machines "sits awkwardly beside his acknowledgment that controlling such godlike entities will be almost impossible."[15] Library Journal called it a "must-read" for technologists, but stated the book was not for the casual reader.[16] The *Wall Street Journal* called it "lucid and engaging"; however, it cautioned readers that the controversial notion that superintelligence could run amok has more credence than it does few years ago, but is still fiercely opposed by many computer scientists.[17]

Rather than endorse a specific future, the book invites readers to think about what future they would like to see, and to discuss their thoughts on the Future of Life Website.[18] The *Wall Street Journal* review called this attitude noble but naive, and criticized the referenced Web site for being "chockablock with promo material for the book".[17]

The hardcover edition was on the general New York Times Best Seller List for two weeks,[19] and made on the New York Times business bestseller list in September and October 2017.[20]

References

1.

Tegmark, Max (2017). *Life 3.0 : being human in the age of artificial intelligence* (First ed.). New York: Knopf. ISBN 9781101946596. OCLC 973137375

. "IBM100 - Deep Blue" . www-03.ibm.com. 2012-03-07. Retrieved 2017-10-20.

Markoff, John (2011-02-16). "On 'Jeopardy!' Watson Win Is All but Trivial" . The New York Times. ISSN 0362-4331 . Retrieved 2017-10-20.

"In Major AI Breakthrough, Google System Secretly Beats Top Player at the Ancient Game of Go" . WIRED. Retrieved 2017-10-20.

Harari, Yuval Noah (2017-09-22). "Life 3.0 by Max Tegmark review – we are ignoring the AI apocalypse" . The Guardian. ISSN 0261-3077 . Retrieved 2017-10-20.

"Podcast: Life 3.0 - Being Human in the Age of Artificial Intelligence - Future of Life Institute" . Future of Life Institute. 2017-08-29. Retrieved 2017-10-20.

https://www.cnbc.com/2017/08/29/elon-musk-recommends-a-book-on-the-future-of-artificial-intelligence.html

Moody, Oliver (30 October 2017). "Why Elon Musk thinks Max Tegmark is the geek who will save the world" . The Times of London. Retrieved 27 November 2017.

LIFE 3.0 by Max Tegmark | Kirkus Reviews .

Russell, Stuart (2017-08-31). "Artificial intelligence: The future is superintelligent" . Nature. **548** (7669): 520–521. doi:10.1038/548520a . ISSN 0028-0836 .

Hirsh, Haym (2017-08-02). "A physicist explores the future of artificial intelligence" . Science Magazine. **357** (6350) (published 2017-08-04). Retrieved 2017-10-19.

Poole, Steven (26 November 2017). "Thinking big, snoozing bigger: the best science books of 2017" . The Telegraph. Retrieved 8 December 2017.

Poole, Steven (2017-08-27). "Artificial intelligence: how scared should we be about machines taking over?" . The Telegraph. ISSN 0307-1235 . Retrieved 2017-10-20.

"3 science books compelling enough to speak to all readers" . Christian Science Monitor. 30 August 2017. Retrieved 11 December 2017.

"Nonfiction Book Review: Life 3.0: Being Human in the Age of Artificial Intelligence by Max Tegmark. Knopf, $28 (384p) ISBN 978-1-101-94659-6" . PublishersWeekly.com. 10 July 2017. Retrieved 7 January 2018.

Browning, Natalie (15 September 2017). "Life 3.0: being human in the age of artificial intelligence". Library Journal.

*Rose, Frank (2017-08-28). "When Machines Run Amok"
. Wall Street Journal. ISSN 0099-9660
. Retrieved 2017-10-20.*

*"Superintelligence survey - Future of Life Institute"
. Future of Life Institute. Retrieved 7 January 2018.*

*"Hardcover Nonfiction Books - Best Sellers - September 24, 2017 - The New York Times"
. 24 September 2017. Retrieved 10 February 2018.*

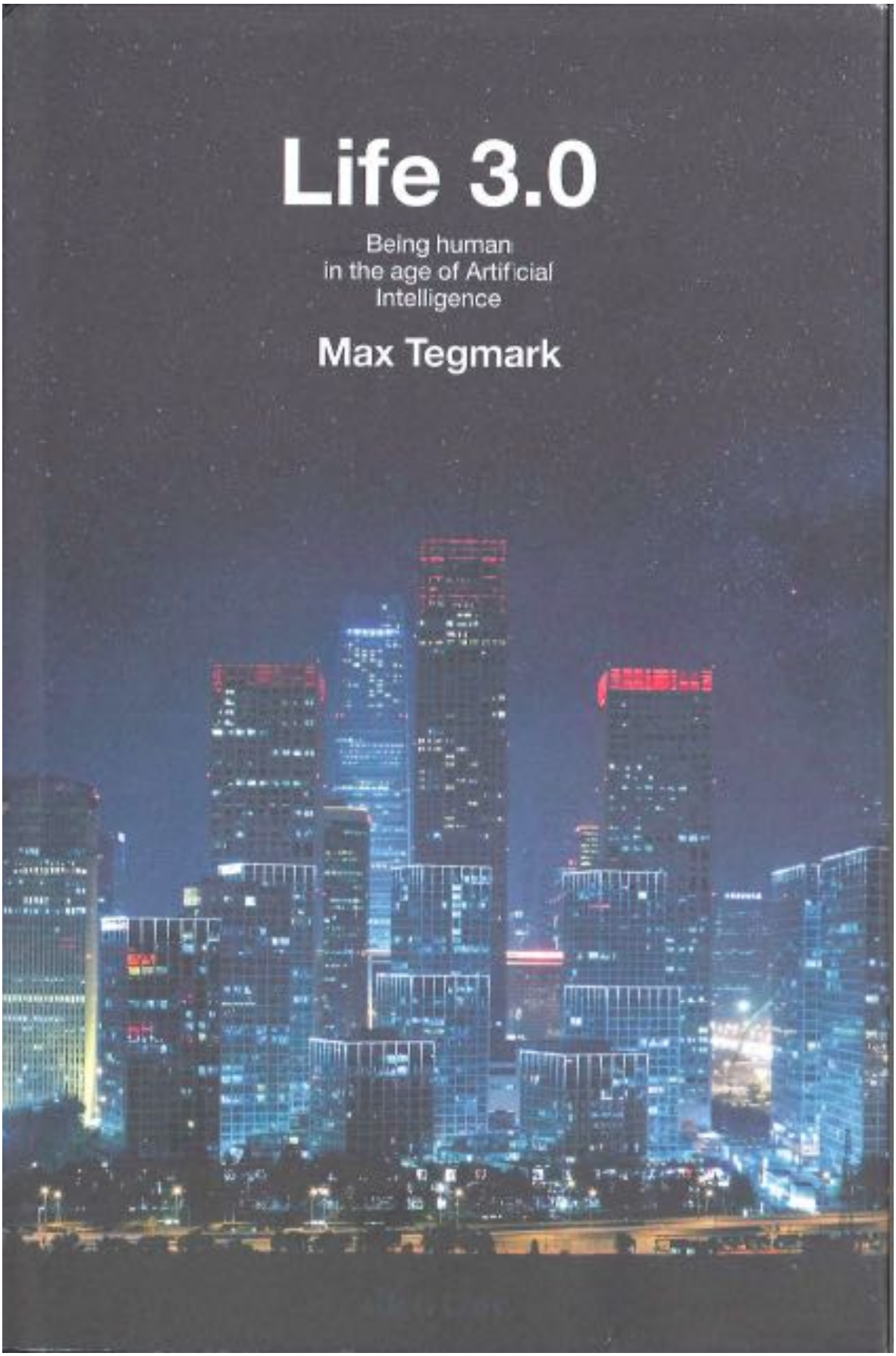
1. *"Business Books - Best Sellers - September 2017 - The New York Times"*

.

External links

- Excerpt from the book
- "Myths and Facts About Superintelligent AI"

on YouTube (a video commissioned by Tegmark's FLI to explain the book)



1.- De quoi exercer les activités de vos cellules grises pendant l'été en pensant à Phil

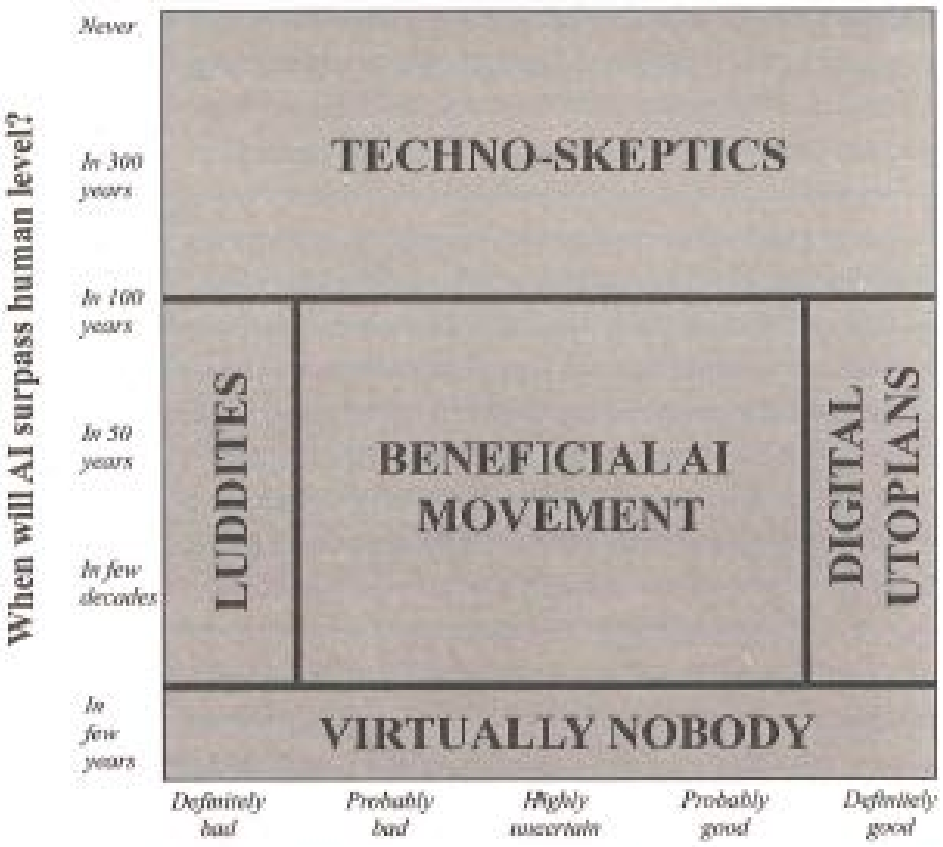


Figure 1.2: Most controversies surrounding strong artificial intelligence (that can match humans on any cognitive task) center around two questions: When (if ever) will it happen, and will it be a good thing for humanity? Techno-skeptics and digital utopians agree that we shouldn't worry, but for very different reasons: the former are convinced that human-level artificial general intelligence (AGI) won't happen in the foreseeable future, while the latter think it will happen but is virtually guaranteed to be a good thing. The beneficial-AI movement feels that concern is warranted and useful, because AI-safety research and discussion now increases the chances of a good outcome. Luddites are convinced of a bad outcome and oppose AI. This figure is partly inspired by Tim Urban.[1]

found literary classic *The Day My Butt Went Psycho*, by Andy Griffiths, and Larry ordered it on the spot. I struggled to remind myself that he might go down in history as the most influential human ever to have lived: my guess is that if superintelligent digital life engulfs our Universe in my lifetime, it will be because of Larry's decisions.

With our wives, Lucy and Meia, we ended up having dinner together and discussing whether machines would necessarily be con-

Terminology Cheat Sheet	
Life	Process that can retain its complexity and replicate
Life 1.0	Life that evolves its hardware and software (biological stage)
Life 2.0	Life that evolves its hardware but designs much of its software (cultural stage)
Life 3.0	Life that designs its hardware and software (technological stage)
Intelligence	Ability to accomplish complex goals
Artificial Intelligence (AI)	Non-biological intelligence
Narrow intelligence	Ability to accomplish a narrow set of goals, e.g., play chess or drive a car
General intelligence	Ability to accomplish virtually any goal, including learning
Universal intelligence	Ability to acquire general intelligence given access to data and resources
[Human-level] Artificial General Intelligence (AGI)	Ability to accomplish any cognitive task at least as well as humans
Human-level AI	AGI
Strong AI	AGI
Superintelligence	General intelligence far beyond human level
Civilization	Interacting group of intelligent life forms
Consciousness	Subjective experience
Qualia	Individual instances of subjective experience
Ethics	Principles that govern how we should behave
Teleology	Explanation of things in terms of their goals or purposes rather than their causes
Goal-oriented behavior	Behavior more easily explained via its effect than via its cause
Having a goal	Exhibiting goal-oriented behavior
Having purpose	Serving goals of one's own or of another entity
Friendly AI	Superintelligence whose goals are aligned with ours
Cyborg	Human-machine hybrid
Intelligence explosion	Recursive self-improvement rapidly leading to superintelligence
Singularity	Intelligence explosion
Universe	The region of space from which light has had time to reach us during the 13.8 billion years since our Big Bang

Table 1.1: Many misunderstandings about AI are caused by people using the words above to mean different things. Here's what I take them to mean in this book. (Some of these definitions will only be properly introduced and explained in later chapters.)

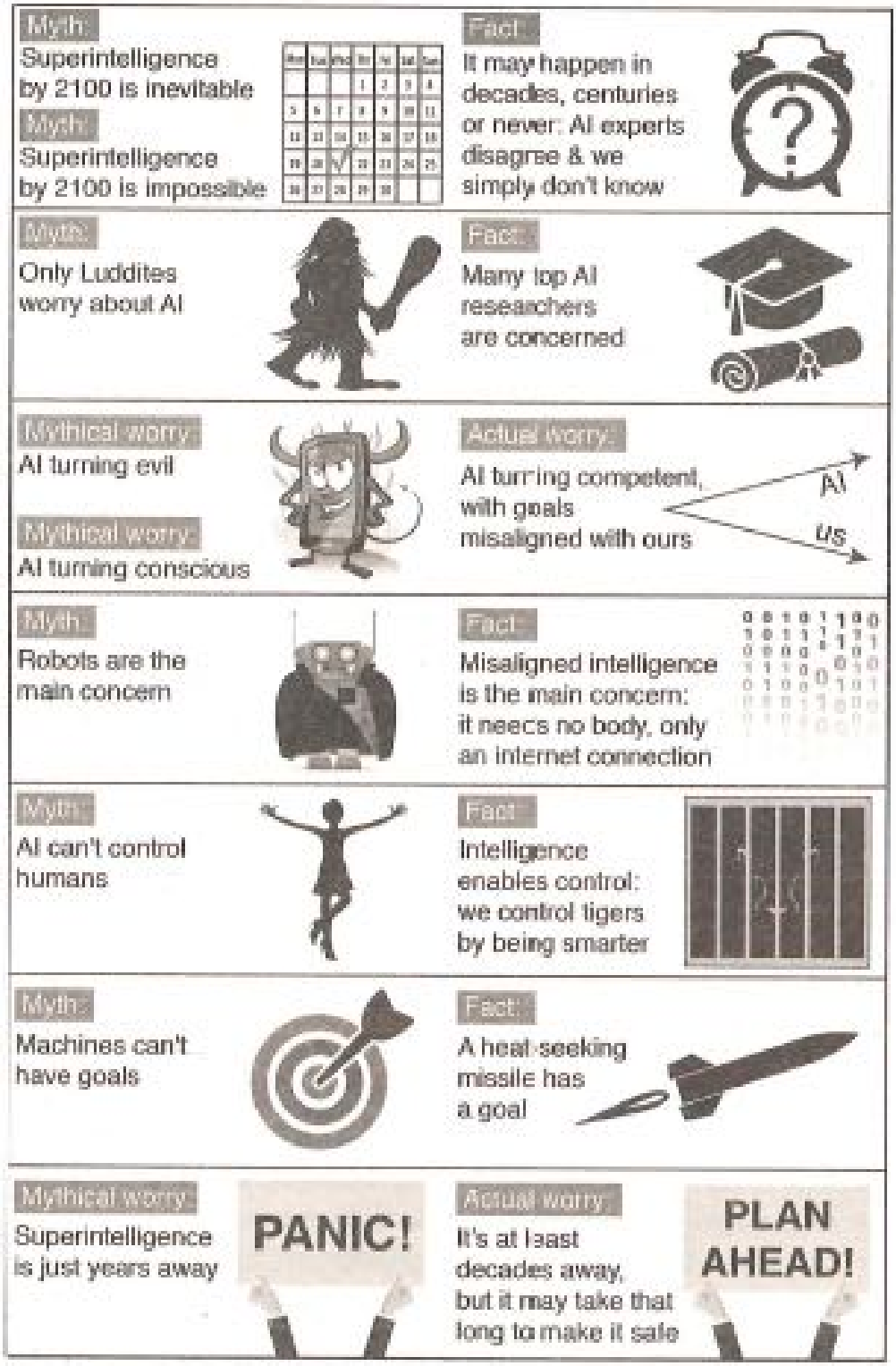


Figure 1.5: Common myths about superintelligent AI.

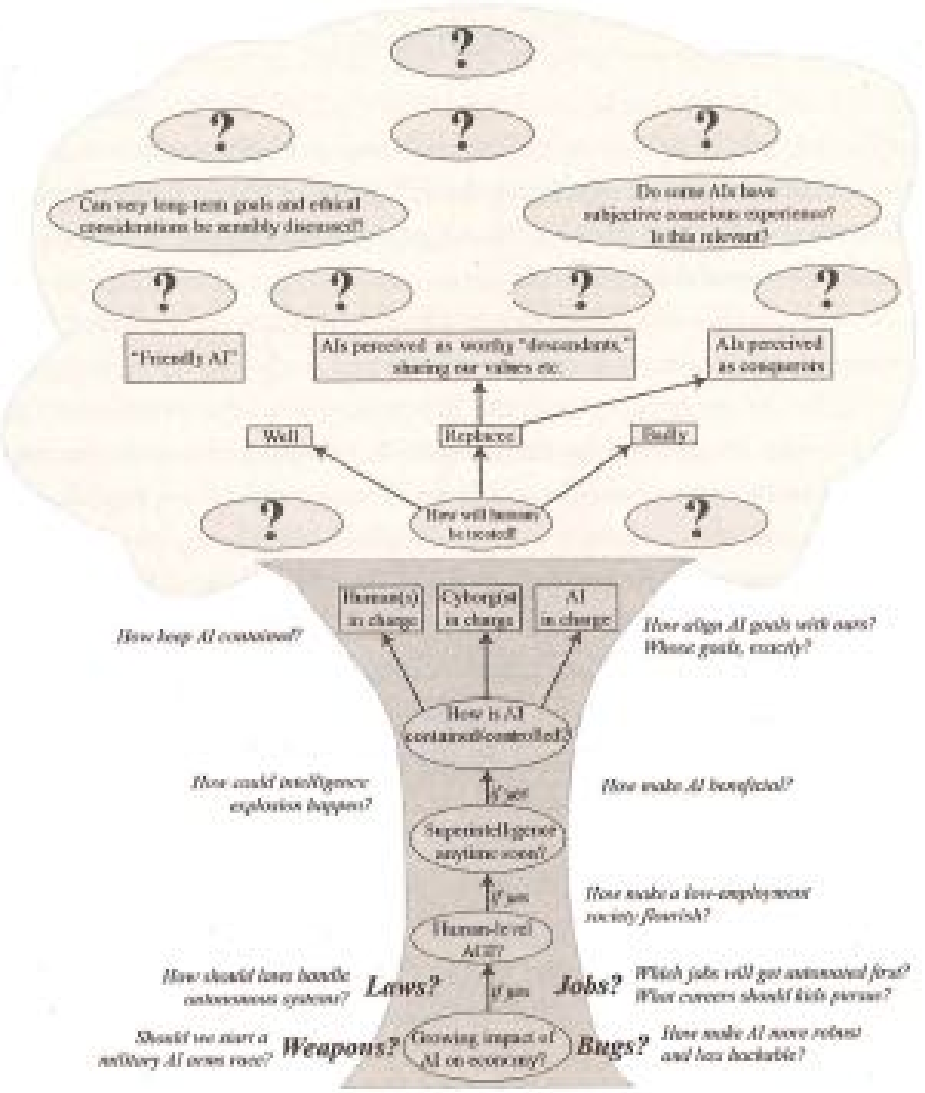


Figure 1.6: Which AI questions are interesting depends on how advanced AI gets and which branch our future takes.

we'll devote the remainder of the book to considering what future to aim for and how to get there. To be able to link cold facts to questions of purpose and meaning, we explore the physical basis of goals in chapter 7 and consciousness in chapter 8. Finally, in the epilogue, we explore what can be done right now to help create the future we want.

In case you're a reader who likes skipping around, most chapters are relatively self-contained once you've digested the terminology and definitions from this first chapter and the beginning of the next

		Short Chapter Title	*Topic*	*Status*
		Prelude: Tale of the Omega Team	Food for thought	Extremely Speculative
The history of intelligence	1	The Conversation	Key ideas, terminology	Not very speculative
	2	Matter Turns Intelligent	Fundamentals of intelligence	
	3	AI, Economics, Weapons & Law	Near future	
	4	Intelligence Explosion?	Superintelligence scenarios	Extremely Speculative
	5	Aftermath	Subsequent 10,000 years	
	6	Our Cosmic Endowment	Subsequent billions of years	
The history of meaning	7	Goals	History of goals-oriented behavior	Not very speculative
	8	Consciousness	Natural & artificial consciousness	Speculative
		Epilogue: Tale of the FLI Team	What should we do?	Not very speculative

Figure 1.7: Structure of the book

one. If you're an AI researcher, you can optionally skip all of chapter 2 except for its initial intelligence definitions. If you're new to AI, chapters 2 and 3 will give you the arguments for why chapters 4 through 6 can't be trivially dismissed as impossible science fiction. Figure 1.7 summarizes where the various chapters fall on the spectrum from factual to speculative.

A fascinating journey awaits us. Let's begin!

THE BOTTOM LINE:

- Life, defined as a process that can retain its complexity and replicate, can develop through three stages: a biological stage (1.0), where its hardware and software are evolved, a cultural stage (2.0), where it can design its software (through learning) and a technological stage (3.0), where it can design its hardware as well, becoming the master of its own destiny.
- Artificial intelligence may enable us to launch Life 3.0 this century, and a fascinating conversation has sprung up regarding what future we should aim for and how this can be accomplished. There are three main camps in the controversy: techno-skeptics, digital utopians and the beneficial-AI movement.
- Techno-skeptics view building superhuman AGI as so hard that it won't happen for hundreds of years, making it silly to worry about it (and Life 3.0) now.
- Digital utopians view it as likely this century and wholeheartedly welcome Life 3.0, viewing it as the natural and desirable next step in the cosmic evolution.
- The beneficial-AI movement also views it as likely this century, but views a good outcome not as guaranteed, but as something that needs to be ensured by hard work in the form of AI-safety research.
- Beyond such legitimate controversies where world-leading experts disagree, there are also boring pseudo-controversies caused by misunderstandings. For example, never waste time arguing about "life," "intelligence," or "consciousness" before ensuring that you and your protagonist are using these words to mean the same thing! This book uses the definitions in table 1.1.
- Also beware the common misconceptions in figure 1.5: "Superintelligence by 2100 is inevitable/impossible." "Only Luddites worry about AI." "The concern is about AI turning evil and/or conscious, and it's just years away." "Robots are the main concern." "AI can't control humans and can't have goals."
- In chapters 2 through 6, we'll explore the story of intelligence from its humble beginning billions of years ago to possible cosmic futures billions of years from now. We'll first investigate near-term challenges such as jobs, AI weapons and the quest for human-level AGI, then explore possibilities for a fascinating spectrum of possible futures with intelligent machines and/or humans. I wonder which options you'll prefer!
- In chapters 7 through 9, we'll switch from cold factual descriptions to an exploration of goals, consciousness and meaning, and investigate what we can do right now to help create the future we want.
- I view this conversation about the future of life with AI as the most important one of our time—please join it!

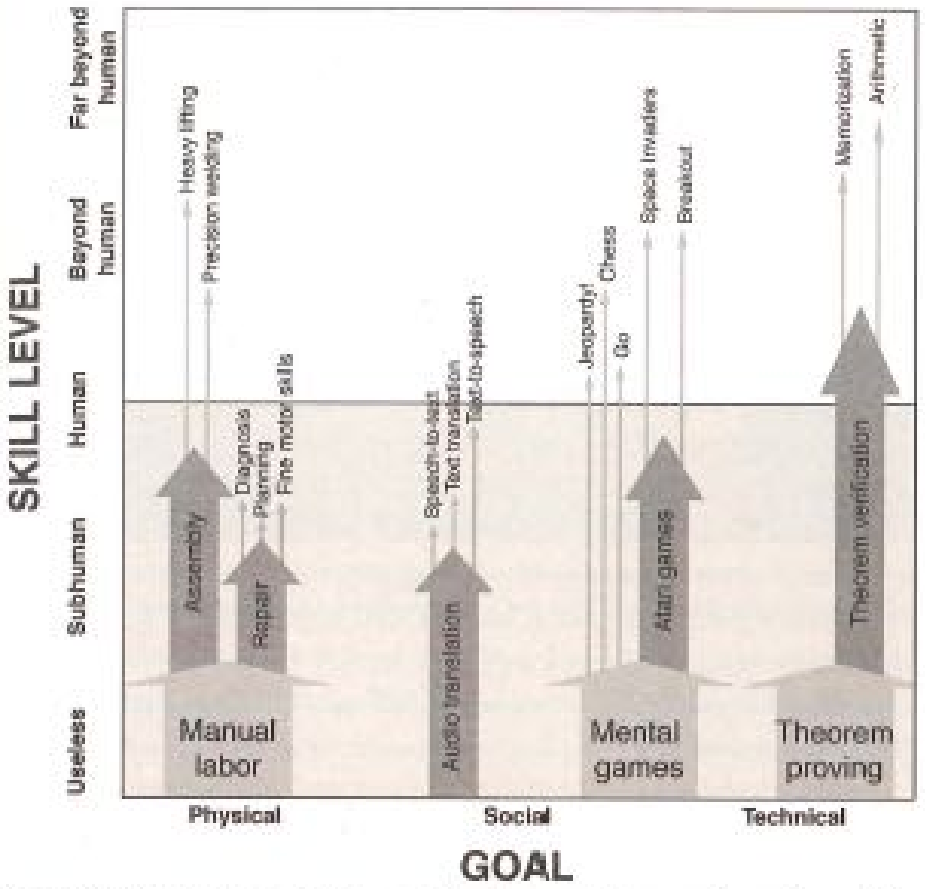


Figure 2.1: Intelligence, defined as ability to accomplish complex goals, can't be measured by a single IQ, only by an ability spectrum across all goals. Each arrow indicates how skilled today's best AI systems are at accomplishing various goals, illustrating that today's artificial intelligence tends to be *narrow*, with each system able to accomplish only very specific goals. In contrast, human intelligence is remarkably broad: a healthy child can learn to get better at almost anything.

ligence and non-intelligence, and it's more useful to simply quantify the degree of ability for accomplishing different goals.

To classify different intelligences into a taxonomy, another crucial distinction is that between *narrow* and *broad* intelligence. IBM's Deep Blue chess computer, which dethroned chess champion Garry Kasparov in 1997, was only able to accomplish the very narrow task of playing chess—despite its impressive hardware and software, it couldn't even beat a four-year-old at tic-tac-toe. The DQN AI system of Google DeepMind can accomplish a slightly broader range

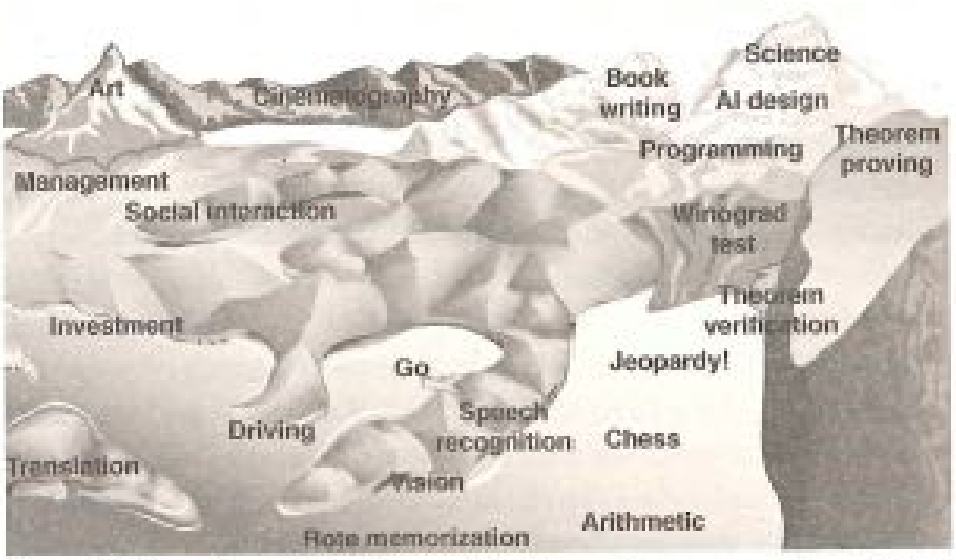


Figure 2.2: Illustration of Hans Moravec's "landscape of human competence," where elevation represents difficulty for computers, and the rising sea level represents what computers are able to do.

it into a hierarchy of subgoals of its own, from paying the cashier to grating the Parmesan. In this sense, intelligent behavior is inexorably linked to goal attainment.

It's natural for us to rate the difficulty of tasks relative to how hard it is for us humans to perform them, as in figure 2.1. But this can give a misleading picture of how hard they are for computers. It feels much harder to multiply 314,159 by 271,828 than to recognize a friend in a photo, yet computers creamed us at arithmetic long before I was born, while human-level image recognition has only recently become possible. This fact that low-level sensorimotor tasks seem easy despite requiring enormous computational resources is known as Moravec's paradox, and is explained by the fact that our brain makes such tasks feel easy by dedicating massive amounts of customized hardware to them—more than a quarter of our brains, in fact.

I love this metaphor from Hans Moravec, and have taken the liberty to illustrate it in figure 2.2:

> Computers are universal machines, their potential extends uniformly over a boundless expanse of tasks. Human potentials, on the other hand, are strong in areas long important for survival, but weak in things far removed. Imagine a "landscape of human com-

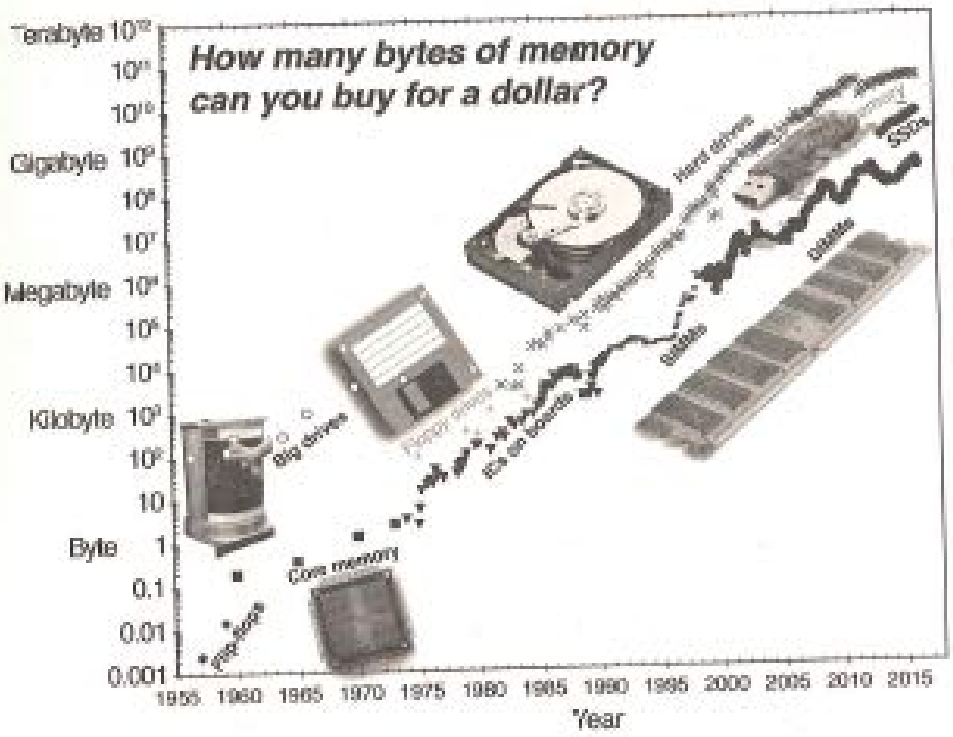


Figure 2.4: Over the past six decades, computer memory has gotten twice as cheap roughly every couple of years, corresponding to a thousand times cheaper roughly every twenty years. A byte equals eight bits. Data courtesy of John McCallum, from http://www.jcmit.net/memoryprice.htm.

book, and then my first-ever hard drive storing 10MB—which might just barely fit a single one of today's song downloads. These memories from my adolescence felt almost unreal the other day, when I spent about $100 on a hard drive with 300,000 times more capacity.

What about memory devices that evolved rather than being designed by humans? Biologists don't yet know what the first-ever life form was that copied its blueprints between generations, but it may have been quite small. A team led by Philipp Holliger at Cambridge University made an RNA molecule in 2016 that encoded 412 bits of generic information and was able to copy RNA strands longer than itself, bolstering the "RNA world" hypothesis that early Earth life involved short self-replicating RNA snippets. So far, the smallest memory device known to be evolved and used in the wild is the genome of the bacterium Candidatus Carsonella ruddii, storing about

late the behavior of quantum-mechanical systems, including atoms, molecules and new materials, replacing measurements in chemistry labs in the same way that simulations on traditional computers have replaced measurements in wind tunnels.

What Is Learning?

Although a pocket calculator can crush me in an arithmetic contest, it will never improve its speed or accuracy, no matter how much it practices. It doesn't learn: for example, every time I press its square-root button, it computes exactly the same function in exactly the same way. Similarly, the first computer program that ever beat me at chess never learned from its mistakes, but merely implemented a function that its clever programmer had designed to compute a good next move. In contrast, when Magnus Carlsen lost his first game of chess at age five, he began a learning process that made him the World Chess Champion eighteen years later.

The ability to learn is arguably the most fascinating aspect of general intelligence. We've already seen how a seemingly dumb clump of matter can remember and compute, but how can it learn? We've seen that finding the answer to a difficult question corresponds to computing a function, and that appropriately arranged matter can calculate any computable function. When we humans first created pocket calculators and chess programs, *we* did the arranging. For matter to learn, it must instead rearrange *itself* to get better and better at computing the desired function—simply by obeying the laws of physics.

To demystify the learning process, let's first consider how a very simple physical system can learn the digits of π and other numbers. Above we saw how a surface with many valleys (see figure 2.3) can be used as a memory device: for example, if the bottom of one of the valleys is at position $x = \pi \approx 3.14159$ and there are no other valleys nearby, then you can put a ball at $x = 3$ and watch the system compute the missing decimals by letting the ball roll down to the bottom. Now, suppose that the surface is made of soft clay and starts out completely flat, as a blank slate. If some math enthusiasts repeatedly place the ball at the locations of each of their favorite numbers, then gravity

late the behavior of quantum-mechanical systems, including atoms, molecules and new materials, replacing measurements in chemistry labs in the same way that simulations on traditional computers have replaced measurements in wind tunnels.

What Is Learning?

Although a pocket calculator can crush me in an arithmetic contest, it will never improve its speed or accuracy, no matter how much it practices. It doesn't learn: for example, every time I press its square-root button, it computes exactly the same function in exactly the same way. Similarly, the first computer program that ever beat me at chess never learned from its mistakes, but merely implemented a function that its clever programmer had designed to compute a good next move. In contrast, when Magnus Carlsen lost his first game of chess at age five, he began a learning process that made him the World Chess Champion eighteen years later.

The ability to learn is arguably the most fascinating aspect of general intelligence. We've already seen how a seemingly dumb clump of matter can remember and compute, but how can it learn? We've seen that finding the answer to a difficult question corresponds to computing a function, and that appropriately arranged matter can calculate any computable function. When we humans first created pocket calculators and chess programs, *we* did the arranging. For matter to learn, it must instead rearrange *itself* to get better and better at computing the desired function—simply by obeying the laws of physics.

To demystify the learning process, let's first consider how a very simple physical system can learn the digits of π and other numbers. Above we saw how a surface with many valleys (see figure 2.3) can be used as a memory device: for example, if the bottom of one of the valleys is at position $x = \pi \approx 3.14159$ and there are no other valleys nearby, then you can put a ball at $x = 3$ and watch the system compute the missing decimals by letting the ball roll down to the bottom. Now, suppose that the surface is made of soft clay and starts out completely flat, as a blank slate. If some math enthusiasts repeatedly place the ball at the locations of each of their favorite numbers, then gravity

THE BOTTOM LINE:

- Intelligence, defined as ability to accomplish complex goals, can't be measured by a single IQ, only by an ability spectrum across all goals.
- Today's artificial intelligence tends to be *narrow*, with each system able to accomplish only very specific goals, while human intelligence is remarkably *broad*.
- Memory, computation, learning and intelligence have an abstract, intangible and ethereal feel to them because they're *substrate-independent*: able to take on a life of their own that doesn't depend on or reflect the details of their underlying material substrate.
- Any chunk of matter can be the substrate for *memory* as long as it has many different stable states.
- Any matter can be *computronium*, the substrate for *computation*, as long as it contains certain universal building blocks that can be combined to implement any function. NAND gates and neurons are two important examples of such universal "computational atoms."
- A neural network is a powerful substrate for *learning* because, simply by obeying the laws of physics, it can rearrange itself to get better and better at implementing desired computations.
- Because of the striking simplicity of the laws of physics, we humans only care about a tiny fraction of all imaginable computational problems, and neural networks tend to be remarkably good at solving precisely this tiny fraction.
- Once technology gets twice as powerful, it can often be used to design and build technology that's twice as powerful in turn, triggering repeated capability doubling in the spirit of Moore's law. The cost of information technology has now halved roughly every two years for about a century, enabling the information age.
- If AI progress continues, then long before AI reaches human level for all skills, it will give us fascinating opportunities and challenges involving issues such as bugs, laws, weapons and jobs—which we'll explore in the next chapter.

THE BOTTOM LINE:

- Near-term AI progress has the potential to greatly improve our lives in myriad ways, from making our personal lives, power grids and financial markets more efficient to saving lives with self-driving cars, surgical bots and AI diagnosis systems.
- When we allow real-world systems to be controlled by AI, it's crucial that we learn to make AI more robust, doing what we want it to do. This boils down to solving tough technical problems related to verification, validation, security and control.
- This need for improved robustness is particularly pressing for AI-controlled weapon systems, where the stakes can be huge.
- Many leading AI researchers and roboticists have called for an international treaty banning certain kinds of autonomous weapons, to avoid an out-of-control arms race that could end up making convenient assassination machines available to everybody with a full wallet and an axe to grind.
- AI can make our legal systems more fair and efficient if we can figure out how to make robojudges transparent and unbiased.
- Our laws need rapid updating to keep up with AI, which poses tough legal questions involving privacy, liability and regulation.
- Long before we need to worry about intelligent machines replacing us altogether, they may increasingly replace us on the job market.
- This need not be a bad thing, as long as society redistributes a fraction of the AI-created wealth to make everyone better off.
- Otherwise, many economists argue, inequality will greatly increase.
- With advance planning, a low-employment society should be able to flourish not only financially, with people getting their sense of purpose from activities other than jobs.
- Career advice for today's kids: Go into professions that machines are bad at—those involving people, unpredictability and creativity.
- There's a non-negligible possibility that AGI progress will proceed to human levels and beyond—we'll explore that in the next chapter!

THE BOTTOM LINE:

- If we one day succeed in building human-level AGI, this may trigger an intelligence explosion, leaving us far behind.
- If a group of humans manage to control an intelligence explosion, they may be able to take over the world in a matter of years.
- If humans fail to control an intelligence explosion, the AI itself may take over the world even faster.
- Whereas a rapid intelligence explosion is likely to lead to a single world power, a slow one dragging on for years or decades may be more likely to lead to a multipolar scenario with a balance of power between a large number of rather independent entities.
- The history of life shows it self-organizing into an ever more complex hierarchy shaped by collaboration, competition and control. Superintelligence is likely to enable coordination on ever-larger cosmic scales, but it's unclear whether it will ultimately lead to more totalitarian top-down control or more individual empowerment.
- Cyborgs and uploads are plausible, but arguably not the fastest route to advanced machine intelligence.
- The climax of our current race toward AI may be either the best or the worst thing ever to happen to humanity, with a fascinating spectrum of possible outcomes that we'll explore in the next chapter.
- We need to start thinking hard about which outcome we prefer and how to steer in that direction, because if we don't know what we want, we're unlikely to get it.

THE BOTTOM LINE:

- The current race toward AGI can end in a fascinatingly broad range of aftermath scenarios for upcoming millennia.
- Superintelligence can peacefully coexist with humans either because it's forced to (enslaved-god scenario) or because it's "friendly AI" that wants to (libertarian-utopia, protector-god, benevolent-dictator and zookeeper scenarios).
- Superintelligence can be prevented by an AI (gatekeeper scenario) or by humans (1984 scenario), by deliberately forgetting the technology (reversion scenario) or by lack of incentives to build it (egalitarian-utopia scenario).
- Humanity can go extinct and get replaced by AIs (conqueror and descendant scenarios) or by nothing (self-destruction scenario).
- There's absolutely no consensus on which, if any, of these scenarios are desirable, and all involve objectionable elements. This makes it all the more important to continue and deepen the conversation around our future goals, so that we don't inadvertently drift or steer in an unfortunate direction.

THE BOTTOM LINE:

- Compared to cosmic timescales of billions of years, an intelligence explosion is a sudden event where technology rapidly plateaus at a level limited only by the laws of physics.
- This technological plateau is vastly higher than today's technology, allowing a given amount of matter to generate about ten billion times more energy (using sphalerons or black holes), store 12–18 orders of magnitude more information or compute 31–41 orders of magnitude faster—or to be converted to any other desired form of matter.
- Superintelligent life would not only make such dramatically more efficient use of its existing resources, but would also be able to grow today's biosphere by about 32 orders of magnitude by acquiring more resources through cosmic settlement at near light speed.
- Dark energy limits the cosmic expansion of superintelligent life and also protects it from distant expanding death bubbles or hostile civilizations. The threat of dark energy tearing cosmic civilizations apart motivates massive cosmic engineering projects, including wormhole construction if this turns out to be feasible.
- The main commodity shared or traded across cosmic distances is likely to be information.
- Barring wormholes, the light-speed limit on communication poses severe challenges for coordination and control across a cosmic civilization. A distant central hub may incentivize its superintelligent "nodes" to cooperate either through rewards or through threats, say by deploying a local guard AI programmed to destroy the node by setting off a supernova or quasar unless the rules are obeyed.
- The collision of two expanding civilizations may result in assimilation, cooperation or war, where the latter is arguably less likely than it is between today's civilizations.
- Despite popular belief to the contrary, it's quite plausible that we're the only life form capable of making our observable Universe come alive in the future.
- If we don't improve our technology, the question isn't whether humanity will go extinct, but merely how: will an asteroid, a supervolcano, the burning heat of the aging Sun or some other calamity get us first?
- If we do keep improving our technology with enough care, foresight and planning to avoid pitfalls, life has the potential to flourish on Earth and far beyond for many billions of years, beyond the wildest dreams of our ancestors.

Figure 7.2: Any ultimate goal of a superintelligent AI naturally leads to the subgoals shown. But there's an inherent tension between goal retention and improving its world model, which casts doubts on whether it will actually retain its original goal as it gets smarter.

The way I see it, the basic argument is that to maximize its chances of accomplishing its ultimate goals, whatever they are, an AI should pursue the subgoals shown in Figure 7.2. It should strive not only to improve its capability of achieving its ultimate goals, but also to ensure that it will retain these goals even after it has become more capable. This sounds quite plausible: After all, would you choose to get an IQ-boosting brain implant if you knew that it would make you want to kill your loved ones? This argument that an ever more intelligent AI will retain its ultimate goals forms a cornerstone of the friendly-AI vision promulgated by Eliezer Yudkowsky and others: it basically says that if we manage to get our self-improving AI to become friendly by learning and adopting our goals, then we're all set, because we're guaranteed that it will try its best to remain friendly forever.

But is it really true? To answer this question, we need to also explore the other emergent subgoals from figure 7.2. The AI will obviously maximize its chances of accomplishing its ultimate goal, whatever it is, if it can enhance its capabilities, and it can do this by

THE BOTTOM LINE:

- The ultimate origin of goal-oriented behavior lies in the laws of physics, which involve optimization.
- Thermodynamics has the built-in goal of *dissipation:* to increase a measure of messiness that's called *entropy*.
- *Life* is a phenomenon that can help dissipate (increase overall messiness) even faster by retaining or growing its complexity and replicating while increasing the messiness of its environment.
- Darwinian evolution shifts the goal-oriented behavior from dissipation to replication.
- Intelligence is the ability to accomplish complex goals.
- Since we humans don't always have the resources to figure out the truly optimal replication strategy, we've evolved useful rules of thumb that guide our decisions: feelings such as hunger, thirst, pain, lust and compassion.
- We therefore no longer have a simple goal such as replication; when our feelings conflict with the goal of our genes, we obey our feelings, as by using birth control.
- We're building increasingly intelligent machines to help us accomplish our goals. Insofar as we build such machines to exhibit goal-oriented behavior, we strive to align the machine goals with ours.
- Aligning machine goals with our own involves three unsolved problems: making machines learn them, adopt them and retain them.
- AI can be created to have virtually any goal, but almost any sufficiently ambitious goal can lead to subgoals of self-preservation, resource acquisition and curiosity to understand the world better—the former two may potentially lead a superintelligent AI to cause problems for humans, and the latter may prevent it from retaining the goals we give it.
- Although many broad ethical principles are agreed upon by most humans, it's unclear how to apply them to other entities, such as non-human animals and future AIs.
- It's unclear how to imbue a superintelligent AI with an ultimate goal that neither is undefined nor leads to the elimination of humanity, making it timely to rekindle research on some of the thorniest issues in philosophy!

THE BOTTOM LINE:

- There's no undisputed definition of "consciousness." I use the broad and non-anthropocentric definition *consciousness = subjective experience.*
- Whether AIs are conscious in that sense is what matters for the thorniest ethical and philosophical problems posed by the rise of AI: Can AIs suffer? Should they have rights? Is uploading a subjective suicide? Could a future cosmos teeming with AIs be the ultimate zombie apocalypse?
- The problem of understanding intelligence shouldn't be conflated with three separate problems of consciousness: the "pretty hard problem" of predicting which physical systems are conscious, the "even harder problem" of predicting qualia, and the "*really* hard problem" of why anything at all is conscious.
- The "pretty hard problem" of consciousness is scientific, since a theory that predicts which of your brain processes are conscious is experimentally testable and falsifiable, while it's currently unclear how science could fully resolve the two harder problems.
- Neuroscience experiments suggest that many behaviors and brain regions are unconscious, with much of our conscious experience representing an after-the-fact summary of vastly larger amounts of unconscious information.
- Generalizing consciousness predictions from brains to machines requires a theory. Consciousness appears to require not a particular kind of particle or field, but a particular kind of information processing that's fairly autonomous and integrated, so that the whole system is rather autonomous but its parts aren't.
- Consciousness might feel so non-physical because it's doubly substrate-independent: if consciousness is the way information feels when being processed in certain complex ways, then it's merely the structure of the information processing that matters, not the structure of the matter doing the information processing.
- If artificial consciousness is possible, then the space of possible AI experiences is likely to be huge compared to what we humans can experience, spanning a vast spectrum of qualia and timescales—all sharing a feeling of having free will.
- Since there can be no meaning without consciousness, it's not our Universe giving meaning to conscious beings, but conscious beings giving meaning to our Universe.
- This suggests that as we humans prepare to be humbled by ever smarter machines, we take comfort mainly in being *Homo sentiens*, not *Homo sapiens*.

THE BOTTOM LINE:

- There's no undisputed definition of "consciousness." I use the broad and non-anthropocentric definition *consciousness = subjective experience.*
- Whether AIs are conscious in that sense is what matters for the thorniest ethical and philosophical problems posed by the rise of AI: Can AIs suffer? Should they have rights? Is uploading a subjective suicide? Could a future cosmos teeming with AIs be the ultimate zombie apocalypse?
- The problem of understanding intelligence shouldn't be conflated with three separate problems of consciousness: the "pretty hard problem" of predicting which physical systems are conscious, the "even harder problem" of predicting qualia, and the "*really* hard problem" of why anything at all is conscious.
- The "pretty hard problem" of consciousness is scientific, since a theory that predicts which of your brain processes are conscious is experimentally testable and falsifiable, while it's currently unclear how science could fully resolve the two harder problems.
- Neuroscience experiments suggest that many behaviors and brain regions are unconscious, with much of our conscious experience representing an after-the-fact summary of vastly larger amounts of unconscious information.
- Generalizing consciousness predictions from brains to machines requires a theory. Consciousness appears to require not a particular kind of particle or field, but a particular kind of information processing that's fairly autonomous and integrated, so that the whole system is rather autonomous but its parts aren't.
- Consciousness might feel so non-physical because it's doubly substrate-independent: if consciousness is the way information feels when being processed in certain complex ways, then it's merely the structure of the information processing that matters, not the structure of the matter doing the information processing.
- If artificial consciousness is possible, then the space of possible AI experiences is likely to be huge compared to what we humans can experience, spanning a vast spectrum of qualia and timescales—all sharing a feeling of having free will.
- Since there can be no meaning without consciousness, it's not our Universe giving meaning to conscious beings, but conscious beings giving meaning to our Universe.
- This suggests that as we humans prepare to be humbled by ever smarter machines, we take comfort mainly in being *Homo sentiens*, not *Homo sapiens*.

Epilogue

The Tale of the FLI Team

The saddest aspect of life right now is that science gathers knowledge faster than society gathers wisdom.

Isaac Asimov

Here we are, my dear reader, at the end of the book, after exploring the origin and fate of intelligence, goals and meaning. So how can we translate these ideas into action? What concretely should we *do* to make our future as good as possible? This is precisely the question I'm asking myself right now as I sit here in my window seat en route from San Francisco back to Boston on January 9, 2017, from the AI conference we just organized in Asilomar, so let me end this book by sharing my thoughts with you.

Meia is catching up on sleep next to me after the many short nights of preparing and organizing. Wow—what a wild week it's been! We managed to bring almost all the people I've mentioned in this book together for a few days to this Puerto Rico sequel, including entrepreneurs such as Elon Musk and Larry Page and AI research leaders from academia and companies such as DeepMind, Google, Facebook, Apple, IBM, Microsoft and Baidu, as well as economists, legal scholars, philosophers and other amazing thinkers (see figure 9.1). The results superseded even my high expectations, and I'm feeling more optimistic about the future of life than I have in a long time. In this epilogue, I'm going to tell you why.

2. Pour en savoir plus

[Life 3.0 - Being human in the Age of Artificial Intelligence by Tegmark Max]

3.- Pour en savoir beaucoup plus Achetez sans délai le bouquin

4. **Lisez-le** ⬅==========

encore plus

https://www.youtube.com/watch?v=Gi8LUnhP5yU

et sur Wikipedia *Life 3.0*

From Wikipedia, the free encyclopedia
Jump to navigation Jump to search

N'oubliez pas ce lien aussi

Life 3.0 by Max Tegmark review – we are ignoring the AI apocalypse by The Guardian