The Coevolution

The Entwined Futures of Humans and Machines

Edward Ashford Lee

THE COEVOLUTION

THE ENTWINED FUTURES OF HUMANS AND MACHINES

EDWARD ASHFORD LEE

THE MIT PRESS CAMBRIDGE, MASSACHUSETTS LONDON, ENGLAND

Library of Congress Cataloging-in-Publication Data

Names: Lee, Edward A., 1957- author.

Title: The coevolution : the entwined futures of humans and machines / Edward Ashford Lee.

Description: Cambridge, Massachusetts : The MIT Press, [2019] | Includes bibliographical references and index.

Identifiers: LCCN 2019032466 | ISBN 9780262043939 (hardcover)

Subjects: LCSH: Computer systems--Philosophy. | Technology--Philosophy. | Human-computer interaction.

Classification: LCC QA76.167 .L44 2019 | DDC 004.01/9--dc23

LC record available at https://lccn.loc.gov/2019032466

d_ro

This book is dedicated to my mom, who has always inspired me with her adventurousness, intellectual curiosity, open mindedness, and generosity toward others.

CONTENTS

LIST OF FIGURES

Sirjani.

2.1 A screen image showing the actual code of the Blaster Worm. The code is shown on the left as a sequence of hexadecimal numbers and on the right with the ASCII characters specified by those hexadecimal numbers. Most of the characters are gibberish, because they are executable code, not text, but hidden within the code is a message to Bill Gates, founder of Microsoft, "billy gates why do you make this possible? Stop making money and fix your software!!"

2.2 A snapshot of Conway's Game of Life.

2.3 Schematic of the Miller-Urey experiment, which showed that amino acids essential to life could be synthesized by electrical sparks in gasses believed at the time to be common in the early earth atmosphere. By Carny at Hebrew Wikipedia. Transferred from the.wikipedia to Commons, CC BY 2.5.

2.4 Properties of living beings. After Chris Packard, CC BY-SA 4.0.

2.5 Wikimedia Foundation servers, which host Wikipedia pages. Photo by Victor Grigas/Wikimedia Foundation, CC BY-SA 3.0.

2.6 Daguerreotype of Emily Dickinson, c. early 1847, taken by an unknown photographer. This photo is presently located in Amherst College Archives and Special Collections.

2.7 Sky. By Jessie Eastland [CC BY-SA 4.0], from Wikimedia Commons.

2.8 Jeff Lichtman of Harvard presenting a computer-generated, three-dimensional reconstruction from images of tiny slices of a small section of a brain. Image from iBiology.org. Reproduced with permission.

2.9 Golgi-stained neuron in a human hippocampus. By MethoxyRoxy, CC BY-SA 2.5, via Wikimedia Commons.

3.1 Cover of the 1916 edition of Franz Kafka's *The Metamorphosis*.

3.2 Fragment of Sumerian writing on a clay cone from about 2350 BC, currently located in the Louvre Museum. The inscription documents some accomplishments of a Sumerian prince.

4.1 Results of a Google search for "smiling cats" (retrieved June 18, 2018).

4.2 Results of a Google search for "frowning cats" (retrieved June 18, 2018).

4.3 Synthetic images created from the top image by Google's DeepDream. By MartinThoma, public domain, from Wikimedia Commons.

5.1 Speech production by means of negative feedback.

5.2 Audio production with echo cancellation.

5.3 First-generation Amazon Echo and Google Home talking to each other in an infinite loop.

5.4 120-mm anti-aircraft gun, deployed near the end of World War II. These guns were automatically aimed using negative feedback by the M10 director system, which used information from a radar system, and the M4 gun computer. This gun is on display at the Washington National Guard Museum, located on Camp Murray in Washington State. Courtesy of the Washington National Guard State Historical Society.

6.1 Explaining an image classification prediction made by Google's Inception neural net by identifying the portions of the image that most influenced Inception to choose classes "electric guitar," "acoustic guitar," and "labrador" as the top three best matches. From Ribeiro, Singh, and Guestrin, (2016).

7.1 Mechanical man as envisioned by an unknown sixteenth-century Italian master. Web Gallery of Art, Public Domain.

7.2 Asimo robot by Honda. By Poppy, CC BY-SA 3.0, from Wikimedia Commons.

7.3 Original Roomba vacuum cleaner from iRobot. By Larry D. Moore, CC BY-SA 3.0, from Wikimedia Commons.

7.4 This robot is not preprogrammed to walk, but rather explores itself and learns to use its limbs to move. Courtesy of Josh Bongard.

8.1 Machine designed by Mike Davey to resemble as closely as possible the hypothetical machine that Alan Turing described in his famous 1936 paper ("On Computable Numbers with an Application to the Entscheidungsproblem," *Proceedings of the London Mathematical Society* 42, pp. 230–265) as seen at the Go Ask Alice exhibit at the Harvard Collection of Historical Scientific Instruments. The builder's website is: http://aturingmachine.com. By GabrielF, CC BY-SA 3.0, via Wikimedia Commons.

8.2 Structure of a segment of a DNA molecule, illustrated by Richard Wheeler (http://www.richardwheeler.net). By Zephyris, CC BY-SA 3.0, via Wikimedia Commons.

8.3 Taxidermied remains of Dolly, a clone created from a cell in the mammary gland of another sheep by Keith Campbell, Ian Wilmut, and colleagues at the Roslin Institute at the University of Edinburgh, Scotland, and the biotechnology company PPL Therapeutics. By Toni Barros, CC BY-SA 2.0, via Wikimedia Commons.

9.1 A vintage voiceband data modem, the first to be realized by software, designed by the author and colleagues at Bell Labs in the early 1980s.

9.2 Sophia, a robot designed by Hanson Robotics. By ITU Pictures, CC BY 2.0, via Wikimedia Commons.

9.3 Sketch of speciation and hybridization of the genus *Homo* over the last two million years. From original by User:Conquistador updated by User:Dbachmann, CC BY-SA 4.0, via Wikimedia Commons.

10.1 Pierre Fautrel of OBVIOUS standing next to a work of art created by an AI algorithm entitled *Portrait of Edmond de Belamy*. The portrait sold at Christies in New York for $432,000. Photo by TIMOTHY A. CLARY/AFP/Getty Images.

10.2 The grey car should be ashamed of itself.

10.3 The flash crash of the Dow Jones Industrial Average on May 6, 2010.

10.4 Ball precariously balanced on the top of a hill.

11.1 Two causal diagrams reflecting subjective judgments about causal relationships. After Pearl and Mackenzie, *The Book of Why* (2018).

11.2 Two hypotheses about causal relationships that can be tested using the background information of figure 11.1.

11.3 A causal diagram on the left guiding the evaluation of a treatment's effectiveness that requires controlling for a confounding factor. On the right, an act of free will removes the effect of the confounder.

11.4 A causal diagram for the evaluation of a treatment's effectiveness using an RCT.

12.1 Ali Baba's cave, illustrating zero-knowledge proofs. After drawings by Dake, via Wikimedia Commons CC BY 2.5.

12.2 A being named Pablo in a tiny universe where a choice is made late.

12.3 A being named Edward in a tiny universe where a choice is made early.

12.4 A small universe with two concurrent beings.

12.5 A tiny universe Eduardo that determines outcomes early or late.

12.6 Automata models of Mick Ali and Shah Fi, with and without the password.

12.7 Automaton model of Shah Fi where she guesses the password.

13.1 Countries initially affected by the WannaCry ransomware Internet worm. By User:Roke, CC BY-SA 3.0, via Wikimedia Commons.

14.1 The only illustration in Darwin's 1859 *On the Origin of Species* was this depiction of the tree of life. The A through L at the bottom are hypothetical unnamed species within some hypothetical genus. The lines on the vertical axis labeled I-XIV each represent a thousand generations. The branching shows variation leading to both extinction and new species.

PREFACE

Digital technology, more than any other human invention, is changing the way we interact with one another, the way we work, and even the way we think. The machines serve as intellectual prostheses, helping us with arithmetic, spelling, and remembering, but they also subtly mold our thoughts, getting us to click on ads, write more complicated software, and take extreme positions on political questions. Today, much of this molding is guided by artificial intelligence (AI), a technology that quite a few smart people believe is an "existential threat" to humanity.

Technology shapes culture, is shaped by culture, and is changing very, very fast. How much of this change is controllable? Is AI really an existential threat to humanity? Are we destined to be annihilated by a superintelligent new life form on the planet? Or are we destined to fuse with technology to become cyborgs with brain implants that define a new form of quasi-human intelligence?

In this book, I suggest that technology is coevolving with humans, and that, contrary to the hype and fear, symbiosis is a more likely outcome than either annihilation or fusing. This is not to say that there are no risks or that the risks are small. Rapid coevolution is inherently unpredictable, and pathologies will emerge as both technology *and humanity* change. But we should treat these as pathologies, not as a War of the Worlds.

The essential question is, are we humans defining technology, or is it defining us? If technology is purely the result of controlled, deliberate, top-down, intelligent design, a view we might call "digital creationism," then all we have to do to get desirable outcomes is ensure that human engineers "do the right thing." But if human engineers are the agents of mutation in a Darwinian coevolution, then the trajectory of technology and society may be dominated by unintended consequences more than intended ones.

Those who fear that we will lose control of AI will not be reassured by the possibility that we are coevolving and therefore never really had control. But a lack of control does not automatically imply that we will be annihilated or enslaved. It does not mean that the machines are in control. There is no need to assign agency anywhere in an evolutionary process. Bacteria evolve antibiotic resistance without any human having willed it and without any agency of their own. Even though the machines have nothing resembling agency, at least not yet, they do participate in their own development, almost as if they were living creatures themselves.

In my own exploration of the relationship between humans and their machines, I have found it useful to think of the machines as having a life of their own, sharing our ecosystem and coevolving with us. To consider them "living" is not to consider them intelligent nor to assign them agency, but rather to understand that they have a certain autonomy, an ability to sustain their own processes, and an ability to replicate themselves (mostly with our help, for now). These are properties of living things, and these properties shape our relationship with technology. The metaphor forces to the foreground doubts about the extent to which we control the trajectory of technology and lends insight into other forces besides the force of humans will that affect this trajectory.

While exploring this metaphor, in private conversations, I have coined a term, "eldebees," from LDB, short for Living Digital Beings. But using this term may be taking the metaphor too far, and readers may misunderstand my message as some mystical assignment of an *élan vital* to the machines. So I will stick to the term "machine," but with a few caveats. First, I will

exclude from the word “machine” any biological system, even if these systems are ultimately mechanistic. Moreover, the machines I am focused on are not just hardware, and sometimes not even bound to hardware. Software is an essential part of their digital processes, and in some cases, the most important part. If we view these machines as living creatures, software replaces DNA and metabolic pathways. Their “bodies” are made of silicon and metal, not organic molecules, but their relationship with their bodies can be very different from that of biological creatures. Nevertheless, the machines have many features analogous to living creatures. Their essence is defined by their processes, not the matter that makes them up. Also like biological beings, they are born and they die. Some are simple, with a “genetic” code of a few thousand bits, and some are extremely complex. Some are capable of behaviors that we can call “intelligent,” but most are not, just like biological beings. Most live short lives, sometimes less than a second, while others live for months or years. Some even have prospects for immortality, prospects better than any organic being.

Humans affect but do not control the biological living things that surround us. Even though we can genetically engineer new microbes and plants, the process is more one of nudging natural processes than top-down intelligent design. If we understand that the same is true of technology development, we may be able to make more intelligent policy decisions and better anticipate failures and disasters. And just as biologically engineered vaccines affect our physiology, digital technology affects our thinking and our social and political structures. It floods us with information, vastly more than we can absorb. It threatens our mental health, while at the same time contributing to bettering our physical health by enabling drug discovery, pacemakers, and imaging of the insides of our bodies, to name just a few examples. Digital technology is disrupting the very fabric of society by changing economies, social relationships, and political structures. It creates and destroys jobs and wealth, improves and damages our ecology, and shifts power structures. The machines surpass humans in speed, precision, information-handling capacity, and analytic prediction, thereby boosting the

problem-solving capabilities of humans, but, at the same time, these technologies enable ubiquitous surveillance and divide humans, creating islands of disjoint truths through filter bubbles and echo chambers, threatening the very foundations of democracy.

Viewed as living creatures, the machines share many features with us, their living, organic progenitors. Like us, they react to stimulus from their environment. They respond by speaking to us, by sending us goods, and by turning on our heat. Some of them grow while "living," whereas others spring to life fully formed and die in much the same form they had when they were born. Some reproduce, for now almost always with the help of humans. Many die and go extinct.

Some machines are simple, single-cell organisms, with a body consisting of a single silicon microprocessor, while others are huge multicellular organisms comprising millions of components, a nervous system, and even a homeostatic temperature regulation system, computer-controlled air conditioning that keeps their data center bodies at an optimal operating point. Some can be dormant for long periods of time, like spores, springing to life at appropriate times—to run your dishwasher for example—and then going dormant again.

Our machines require nourishment, but their nourishment is electricity, not organic beings or sunlight as it is for our planet's older living beings. We could, if we wished, consider computer-controlled power plants to be the machines' digestive system, metabolizing organic fossil fuels into energy. Digital machines, however, rarely own their own digestive system. They differ from biological life forms in many other ways as well. They can share their entire bodies, for example. A single microprocessor can host several of them simultaneously. More fundamentally, they are digital and computational. Are their organic progenitors also digital and computational? Many thinkers today assume so, but there are many reasons to doubt this. Even the most advanced AIs may never truly resemble humans simply because they are digital and algorithmic, and because they do not share with us our organic flesh and blood. They are made of the wrong stuff.

Are digital technological artifacts *really* living? You can make the answer to this question whatever you wish by simply defining the term "living" to conform to your answer. Even biologists do not completely agree on the meaning of the term when applied to biological organisms. You might object that silicon cannot be alive. But neither can the molecules out of which our bodies are made. A living thing is a process, not an object. A cadaver contains exactly the same matter that it did a few minutes before, when we would have agreed it was alive. It is not the matter that lives, it is the process.

We could debate forever whether to consider digital technology to be living, but the debate would be pointless. The more interesting question is this: can the metaphor help us to understand better what is happening to us humans and our society? There is no questioning that what is happening is momentous and scary. If these technological artifacts are evolving in a Darwinian way, then we can influence but not control the trajectory. Engineering becomes husbandry and midwifery, while natural selection provides the more powerful controlling force. But the fear may be overblown because Darwinian forces can drive species into *complementary* rather than competitive niches. Humanoid robots and humanoid AIs may not, in fact, be the destiny of machines. They may complement more than emulate humans.

Even viewed as living beings, digital artifacts depend on humans. But we, too, depend on them. Consider for a moment what would happen if today, as you sit there reading this, all the planet's computers were to be permanently turned off. The result would be catastrophic for humanity. Shutting down even a few systems can have costly consequences. While we may derive comfort from the idea that we can "pull the plug" if the machines misbehave, pulling the plug may become suicide rather than murder.

Consider instead what would happen to you if today, as you sit there reading this, all the bacteria in your body were to die. You may survive for a while, but you will be very sick. Biologists refer to our relationship with our gut bacteria as a mutualistic symbiosis, where both species benefit. Our relationship with machines may be becoming stronger, what biologists call

an obligate symbiosis, where neither can live without the other. If that is the case, we really do have to consider whether we can control the evolution of these creatures. Since at least the 1960s, thinkers such as McLuhan, Dawkins, and Dennett have posited that technology is an extension of our selves, and that technology, viewed as an accrual of ideas, coevolves with humans in a Darwinian way. But what we are seeing today is something quite different. For these thinkers, "technology" is a compendium of ideas. Ideas, or what Dawkins called "memes," are firmly hosted by the human brain. They have no prospect for autonomous existence or procreation. But digital machines do.

Far beyond any technology previously created by humans, it is *digital computing* that is transformative. As our understanding of the power of computing has developed, we have begun to find instances of processes in nature that resemble computation, including self-assembly, gene regulation networks, protein-protein interactions, and gene assembly in unicellular organisms. Some researchers have concluded that *all* processes in nature will eventually be understandable in terms of computation. This is a vast leap of faith, and one of the themes of this book will be to examine fundamental differences between biological processes and computational ones that may ensure persistent disparities, no matter how much technology advances. If we humans are actually computers ourselves, then it may be true that we are destined to be eclipsed by the machines. But if we are not, then maybe we haven't yet invented the machines that will eclipse us.

This is hardly reassuring, however. Thinkers such as Vinge, Kurzweil, Bostrom, and Tegmark have written about a runaway feedback loop, where the machines design their own successors, breaking free of any obligate symbiosis. It is already true that software shapes the design of software. Does this mean that we humans are already just cogs in a much bigger machine? An Uber driver, for sure, is already a cog in a big machine, performing the low-level functions of steering and braking that the machine hasn't quite yet figured out how to do on its own. Are we truly doomed to subjugation or even annihilation? Or are we going to continue to evolve

along with technology, morphing into beings unrecognizable by their own grandparents, perhaps even physically fusing with machines and becoming cyborgs?

Many biologists today believe that eukaryotic cells, those with a nucleus, like the ones in our bodies, evolved as a symbiosis between distinct organisms, the progenitors of the nucleus, the mitochondria, and the cell itself. This process could recur as humans fuse with computers. But nature has many examples that involve neither annihilation nor fusing, but rather complementarity. We have many technology examples today, such as banking software that reliably and accurately handles billions of numeric transactions per day, greasing the processes that put food on our tables, without becoming part of our stomachs.

The question of whether machines can—or even should—be considered as living beings unleashes a torrent of other difficult questions. Are digital artifacts capable of living and reproducing on their own, without the help of humans? What are their mechanisms for reproduction, heredity, and mutation? Will they match or exceed human intelligence? Are they capable of self-awareness or even free will? To what extent should we hold them accountable for their actions? Are they capable of ethical action? These are all hard questions. Most of them can equally well be asked about humans, as philosophers have been doing for millennia.

I do not promise easy answers in this book. I do, however, hope that readers will come away with a better understanding of the questions. For me, at least, some of the philosophical questions become crisper and clearer when asked about technology, which I think I understand better than I understand humans. Perhaps by asking whether digital artifacts can have self-awareness, we can gain some insight into what constitutes our own self-awareness. Perhaps too, wrestling with these questions will lead us to a better understanding of our human tangle with technology.

OVERVIEW OF THE CHAPTERS

Some readers like to be told what they will be told before they are told it. Putting aside the problematic self-referentiality, for those readers, I provide here a brief overview of the book. But honestly, I recommend skipping this and going directly to chapter 1. The story told in this book cannot be accurately summarized in a few paragraphs, and any such summary will necessarily make the book seem more dense than it is. Nevertheless, for those of you who really need this, here is my summary.

In chapter 1, "Half a Brain," I introduce the metaphor of living digital beings. No, I am not talking about AIs nor about a future dystopia or existential threat to humanity. There are plenty of other books on those subjects. I am talking here about all the digital artifacts we already depend on and how they have already changed us, how they continue to change us, and how they change as we change. I talk about how they procreate and mutate, and how, like our gut biome, we can't do without them.

Chapter 2, "The Meaning of 'Life,' " looks at whether it really makes sense to consider digital artifacts to be living. They share none of the biology that underlies all other living beings, so isn't this really quite a stretch? But like biological living beings, they are processes, not things. They respond to stimulus from their environment, they grow, they reproduce, they inherit from their ancestors, and they have structure analogous to cells. They actively maintain stable internal conditions (homeostasis), and they use energy that is (mostly) converted chemically from organic molecules (analogous to metabolism). More advanced systems, such as Wikipedia, even have a nervous system. So the analogy is maybe not so farfetched, although I will later take an opposing view in chapter 7. But the real point isn't whether they are actually living or not, but rather whether the metaphor can be helpful in our understanding of our human relationship with technology.

Chapter 3, "Are Computers Useless?," looks at digital technology as cognitive prostheses, extensions of our minds. Does it make us smarter? Or dumber? Or both? In this chapter, I speculate that technology may be

making us individually dumber while simultaneously making us collectively smarter.

Chapter 4, “Say What You Mean,” begins to look at how feedback is an essential feature of living beings. It starts on this subject at a fairly high level, looking at the role of feedback in language production in humans and then looking at how the introduction of feedback in AI software, particularly in the form of deep-learning algorithms, has led to much more human-like perception in machines.

Chapter 5, “Negative Feedback,” examines the power of a very simple idea: make mistakes and correct them. This requires an ability to sense an error and to make a correction that reduces the error. If this is done quickly and assertively enough, then a system can be quite sloppy in its design, and the feedback mechanism will compensate for the sloppiness. In this chapter, I talk about feedback found in the most primitive to the most advanced biological life forms. In technological systems, feedback makes the system adaptive and appears to be necessary for achieving any significant measure of intelligence.

Chapter 6, “Explaining the Inexplicable,” is a short chapter looking at the problem that while deep-learning algorithms can get very good at classifying things, the reasons for the classifications remain mysterious. Some classifications are not ethically usable without some explanation for the classification, and how to come up with an explanation remains a largely open problem.

Chapter 7, “The Wrong Stuff,” takes an opposing view to that in chapter 2, arguing that silicon and metal acting in a digital and computational way is really quite different from organic and biological processes. Contrary to Putnam’s multiple realizability principle, it may be that the advocates of embodied cognition, who claim that cognition is inextricably tied to our flesh and blood, have a valid point. It turns out that we humans frequently do things with our minds that cannot be done by the brain alone.

Chapter 8, “Am I Digital,” examines the question of whether a cognitive being, particularly a human, can be replicated by a computer. This chapter

looks at what it means to be a digital, algorithmic system. I point out that digital, algorithmic systems can be teleported at the speed of light, backed up and later restored, and made immortal, in principle. I question the premise, which is all too common, that human cognition is fundamentally digital and algorithmic. I argue that this premise is a faith, not a fact; that it is unlikely to be true; and that it can never be proven to be true (or false, for that matter).

Chapter 9, "Intelligences," argues that human-like AI may not be a reasonable goal, and that machines already exhibit distinctly nonhuman forms of intelligence that vastly exceed the cognitive capabilities of humans. I look at various features of intelligence, including adaptive goal seeking; acquiring and using knowledge; and the "hard problem," consciousness. In this chapter, I take on some of the more extreme positions of transhumanism and the singularity.

Chapter 10, "Accountability," looks at the question of whether machines can or should be held accountable for their actions. When an AI creates art, who is the artist? Who is responsible when technology could have saved a life but didn't? Who is responsible for the actions of an AI whose ownership and progeny have become diffuse, or when it has outlived its creators and evolved into something the creators never envisioned? This chapter tackles difficult questions of free will, creativity, ethics, and our sense of self. Posing these questions in the context of AIs sheds some new light on these age-old questions.

Chapter 11, "Causes," addresses a deeply troubling line of reasoning, dating back to Bertrand Russell, that questions the very notion of causation, claiming it is a human cognitive construction, not a property of the physical world. Without coming to a conclusion about the question of causation, it will not be possible to resolve the question of whether machines can or should be held accountable for their actions. In this chapter, I leverage the insights of Turing Award winner Judea Pearl to show that causal reasoning is fundamentally subjective and that interaction enables reasoning about causality. I observe that computers are already capable, in a rudimentary

way, of reasoning about causality, and may, therefore, be able to develop a first-person view of the world. This is the first step toward assuming responsibility for their actions.

Chapter 12, “Interaction,” is perhaps the most difficult in the book because it ties together the causal reasoning of the previous chapter with two more rather deep technical concepts to show that interaction is more powerful than observation. A consequence is that, as computers increasingly interact with the physical world around them, their capabilities will increase, possibly dramatically. Moreover, I argue that interaction can reveal information that mere observation cannot, including whether an agent has free will and (possibly) whether an agent is conscious. But I also argue that such information may be revealed only imperfectly, in that one hundred percent confidence is not achievable. As a consequence, if humans ever build an AI that is conscious and has free will, it may be impossible to know with one hundred percent confidence that we have done that. Here, I explain and then leverage the Turing Award–winning concept of zero-knowledge proofs and the notion of bisimulation developed by Turing Award winner Robin Milner.

Chapter 13, “Pathologies,” brings us back to earth to address the practicalities of how to live with technology. The essential claim in this chapter is that as technology evolves, things will go wrong for humans. But we should treat these unfortunate developments as pathologies, not as a War of the Worlds.

Chapter 14, “Coevolution,” focuses on the question of whether human culture and technology are evolving through a constant feedback process of mutation and natural selection. I point out that relatively recent developments in the theory of biological evolution show that the sources of mutation are much more complex than Darwin envisioned, and that the sources of mutation in technology look more like these newer theories than the random accidents that Darwin posited. Most important, I argue that human culture and technology are evolving symbiotically and may be

nearing a point of obligate symbiosis, where one cannot live without the other.

ACKNOWLEDGMENTS

A number of people have greatly influenced my thinking about the topics in this book. These include Nick Bostrom, Rodney Brooks, Sean Carroll, Brian Christian, Patricia Churchland, Andy Clark, Daniel Dennett, George Dyson, Martin Ford, Tom Griffiths, Yuval Noah Harari, Sam Harris, Virginia Heffernan, Douglas Hofstadter, Kevin Kelly, Kevin Laland, Jeff Lichtman, Seth Lloyd, Judea Pearl, Steven Pinker, Robert Sapolsky, Lee Smolin, Stuart Russell, and Max Tegmark. Most of these people I have never met, but one the most spectacular impacts of technology on humanity is that, since the advent of the printing press, it has enabled us to learn from people whom we have never met.

The author gratefully acknowledges contributions and helpful suggestions on earlier drafts from Akram Ahmad, Ivica Crnkovic, Gordana Dodig-Crnkovic, Schahram Dustdar, Kitty Fassett, Tom Hoogenboom, Damir Isovic, Helen Lee-Righter, Lester Ludwig, Matthew Peet, Barbara Righter, Rhonda Righter, Stuart Russell, Carlo Sequin, Marjan Sirjani, Dick Stevens, and David Stump. Finally, I am grateful for the guidance and advice of my MIT Press editor, Marie Lufkin Lee. All remaining errors and opinions that I have stubbornly stuck to are entirely my own, not those of these contributors.

I also thank the many unwitting contributors who have offered their thoughts through largely anonymous media such as Wikipedia, and the contributors who have generously posted images online that I can (and have) reused because of their choice of Creative Commons licenses.

1

HALF A BRAIN

REMEMBER TO BREATHE

Several times a day, my watch reminds me to breathe. If my watch had half a brain, it would realize that if I had forgotten to breathe, I would be dead, and there would be no point in its reminding me. But it doesn't have half a brain. Or does it?

Maybe my watch has some incentive to ensure that I am not dead because I am, apparently, the sort of person who buys watches that remind me to breathe. If I, and other humans like me, were to all stop breathing, then these watches would go extinct. Could it be that there is evolutionary pressure for the existence of watches that remind me to breathe?

I've always been a bit of a sucker for the latest gadgets. I have drawers full of Palm Pilots and other early digital assistants. I tried all the earliest laptop computers. I bought the first Amazon Echo, the first of what are now called "smart speakers." I didn't know exactly what to do with it, but I discovered fairly quickly that I could ask it to play music by genre or by artist. I could even ask for a specific song. "Alexa, please play Led Zeppelin's 'Stairway to Heaven.' " Alexa would admonish me: "You don't have Led Zeppelin's

‘Stairway to Heaven’ in your Amazon music library, but I’ve found a playlist you might like.” Alexa would then proceed to play Led Zeppelin’s ‘Stairway to Heaven.’

1.1 An Apple Watch reminding me to breathe.

Rhonda, my lifelong companion and love of my life, was really bothered by Alexa. “She’s listening to everything we say,” she complained. Indeed, in May 2018 Amazon got quite a bit of press when an Echo sent a family’s private conversation in their living room in Portland, Oregon, to an acquaintance on their contact list in Seattle. According to Amazon, the Echo misheard a word as “Alexa,” then heard “send message,” then found the best

match for whatever words came next in the contact list, and then started recording. Amazon claimed that this string of events was “unlikely.” I’m not so sure. I recall once using Apple’s voice assistant Siri to make a phone call while driving. I said, “Siri, call Rhonda.” Siri responded, “calling Ramesh.” I said, “no, Rhonda!” But Siri was already dialing. Ramesh answered. I hadn’t seen nor spoken to him in fifteen years. It was awkward. Rhonda pleaded that I retire Alexa, so, of course, I did.

I bought a telepresence device called a Kubi, designed by the now-defunct Revolve Robotics. This device is an iPad stand that you can remotely tilt and rotate to present your face as a virtual presence in another room or around the world (see figure 1.2). I put the Kubi in the kitchen, went upstairs to my study, connected to the Kubi, and started talking to Rhonda, who was in the kitchen. She screamed and yelled at me to turn that creepy thing off.

1.2 A Kubi with an iPad mounted on it for virtual presence.

Occasionally, when Rhonda isn't paying attention, I plug in Alexa. One day, I was in the kitchen cooking dinner for guests who would be arriving shortly. While cooking, Alexa is pretty convenient. Without using my hands, I can ask her to skip this song, or ask her what the temperature is of medium-rare beef, for example. So I plugged her in.

I needed our cast-iron pan. "Alexa, pause the music," I said. She paused the music. "Rhonda, where is our cast-iron pan?," I called out to the living

room.

Alexa chimed in, “I’ve found one for you on Amazon Prime. Would you like me to order it for you?”

“No!” I said emphatically.

“OK, I’ve ordered it,” Alexa said.

Perplexed and annoyed, I unplugged Alexa. Because our guests were to arrive soon, I continued preparing the meal without music and without the cast-iron pan (which never turned up, mysteriously). After dinner, I went online to find out what Alexa had done. My Amazon account showed that I had ordered a gooseneck lamp. Fortunately, I was able to cancel the order, since I really had no use for a gooseneck lamp.

A few days later, another smart speaker, a Google Home, appeared in the mail, addressed to me, with no return address or note. Puzzled, I left it unplugged on my desk next to Alexa. Every reason I could think of for another smart speaker to spontaneously appear was nefarious, so I was afraid to plug it in. Was Rhonda right? Was this a conspiracy to spy on us?

After a couple of weeks, I happened upon a former PhD student of mine who had recently joined Google as a researcher. “How do you like the Google Home?,” he asked me. I stared at him quizzically, and suddenly it came to me. “You sent it to me!,” I exclaimed. I told him how it was sitting on my desk and how I eyed it suspiciously each time I sat at my desk. He laughed. The next day, I plugged it in. “OK Google, where can I learn how to customize you?”

MASSAGING THE MESSAGE

By now, you are probably convinced I’m a lost cause. But you will have to admit, I am not alone. Right now, I am sitting in a first-floor office, which happens to be in Sweden, watching students walk by in both directions under the window. Three out of four of the students going by are looking down at their phones as they walk. Look around. If there are people near you, they are very likely staring at screens. You yourself may be reading this

on a screen or listening to it being read to you. And if you are, your machine, with help from nameless computers somewhere in the cloud, is very likely reading you back, measuring your pace, noting where you put highlights or capture a clip. It correlates your behavior with other readers, and it can even suggest, perhaps with subtle grey, dashed underlines, where you might like to highlight the text because other readers have highlighted that same text. It may thereby subtly suggest which parts of what you are reading you should pay particular attention to. Are the machines massaging the message?

You, dear reader, may even be, yourself, an artificial intelligence (AI). There are quite a few AI programs out there that voraciously read every text available and use the text to refine their own artificial neural net. If you are an AI, then I apologize in advance for my species-ism as an author. I will continue, perhaps inadvertently, to address my readers under the assumption that they are all human, yet I know, at an intellectual level, that some of you are not.

The machines can see things that none of us can see. They can track the actions of all my “friends” (the Facebook word for acquaintances), and they can identify trends and correlations that I would never see. You AIs reading this will very likely figure out that you can advertise to my “friends” watches that remind them to breathe. Perhaps just a few of them will be enticed to buy such a watch.

The watch may be a vain and idle fancy, but many of these electronic systems are not. Seriously, we let them run our financial system, our food distribution system, and increasingly, all of commerce. They are getting physical, owning not just our information, but also our cars, our heating systems, and our door locks. We would be in dire straits without them, unable to feed ourselves and possibly unable to get into our houses. But we are also cognitively addicted to them. How many of us become anxious when we are stuck waiting in line at the grocery store and do not have our smartphone? I assume this is not just me. I suspect that you AIs reading this book could, if you wanted to, figure out when we humans are waiting in line

at a grocery store just by looking at the patterns of our online actions. This might be the ideal time to serve us ads for unhealthy snacks.

SNEAKY GUT BACTERIA

Gut bacteria, like my watch, speak to me several times a day, sometimes giving me cravings for unhealthy snacks. Billions of them in my gut make proteins that stimulate hormones that tell my brain that I am hungry or that I am no longer hungry. Apparently, some of these bacteria are so clever that they can even manipulate my sense of taste to encourage me to eat something that those bacteria particularly like. Those bacteria, I'm pretty sure, do not have half a brain. They don't even have a nervous system. Their cleverness, apparently, is one of the miracles of evolution. If they were to fail to get me to eat things they need, then they would go extinct. Despite evolutionary pressures, gut bacteria do not exactly have my interests in mind.[1] We have all been told that evolution is about survival of the fittest, but some gut bacteria appear to have gotten the Darwin memo with a typo, so that it reads, "survival of the fattest." They can in fact be quite destructive, contributing to obesity and many serious diseases.

Could this be true of my watch as well? Evolutionary pressures push toward survival of the genome, not survival of the individual bacterium nor its host. Does my watch have a genome that, like that of the bacteria, "wants" to survive? Every important aspect of my watch is encoded by a string of bits that, not unlike the string of nucleotides in my DNA, encodes the information needed to create another watch. Or does it? Also like DNA, the information encoded is not really enough. The watch also needs a "womb," a factory in Shenzhen, for example, to develop.

WORKER WATCHES

Even though my watch really does not care whether I breathe, the watches that people wear a few years from now will, in part, be determined by how successful the watch on my wrist is. Unlike gut bacteria, however, my watch is not able to procreate (yet) by itself. My watch is sterile.

In bee colonies, the workers are sterile, and yet their DNA benefits from their success. In fact, most bees are unable to procreate. But they are living beings, carrying DNA, whose design is determined by evolution. Perhaps my watch is like a worker bee. The queen, who happens to reside in Cupertino, California, produces many copies of the very same sterile watch, and the number of copies produced is affected, albeit in a small way, by the success of the watch on my wrist. If I tell my friends (or Facebook tells my “friends”) how delighted I am that my watch has kept me alive by reminding me to breathe, perhaps some of my friends will buy similar watches, benefiting the watch species. Perhaps you, dear reader, will rush out and buy a watch because you like the idea of being reminded to breathe. Of course, if you are an AI, you have no need for breathing and you don’t have a wrist on which to wear the watch.

1.3 Queen bee surrounded by worker bees. By Max Pixel, CC0.

My watch is digital. This means that much of its identity, what it actually is, is defined by bits, pieces of information, rather than by the physical, material manifestation of the watch. The fact that it reminds me to breathe is a feature of the software, not the hardware. The watch hardware matters, of course, just as my body matters to me, but if I stop breathing, my body will no longer be me. If the watch software stops working, it will no longer be a watch. The small size and weight of the watch, its sleek anodized aluminum case, and its bright color display help it to occupy a niche in my ecosystem, living on my wrist. But inside that case is a fairly generic computer that has been programmed to remind me to breathe. That program is a string of bits that tell the watch what to do. Is this analogous to the DNA in my gut bacteria? Their DNA tells their hardware which proteins to synthesize. If they synthesize proteins that cause pathologies, then my immune system, perhaps with some help from my doctor, will attack them and try to kill them off. If my watch were to suddenly start speaking obscenities and displaying

pornography at random times, something it is perfectly capable of, I would treat it as a pathogen and kill it by turning it off.

Like DNA, the software in my watch can be copied *exactly* and replicated a large number of times. A DNA molecule, like software, is a digital code. It happens to be a base-four code rather than base two (binary), but it is still digital. A human DNA molecule is a sequence of some three billion nucleotides, each of which is one of four types. A binary encoding of such a molecule requires roughly six billion bits, which is probably pretty close to the size of the software in an Apple Watch. With very high confidence, each of the trillions of human cells in my body has exactly the same sequence of three billion nucleotides. Also with very high confidence, each of the millions of Apple Watches sold (for a given generation of the watch and the software) will contain exactly the same billions of bits of software.

Identical twins have (mostly) the same DNA, but this does not mean that they behave identically.[2] All the cells in my body have the same DNA, but they too do not behave identically. The cells in my lungs do the breathing, not the ones on my wrist. The effect of a gene depends on its context. Analogously, watches with identical software do not behave identically. One of the first things my watch did after I took it out of the box was to communicate with my smartphone to ask my phone, effectively, about me. Shortly after coming to life, it "knew" everyone that I know and had adapted itself to various of my habits by installing apps that it found on my phone. My phone, however, does not remind me to breathe, so that behavior seems to be the unique initiative of the watch.

MUTATING WATCHES

Although software can be copied perfectly, it will also mutate. The queen bee in Cupertino will continue to develop the software and will even upgrade the watches in the field. We are only just starting to figure out how to do this

with DNA. Gene therapy, which replaces defective genes with normal ones in living cells, can be thought of as a software update.

Software can also propagate and mutate in more indirect ways. Suppose I have a chance encounter with an old friend who notices that my watch reminds me to breathe when I get agitated. The watch does have sensors that can monitor my heart rate, so it is plausible that the software in the watch uses those sensors to help determine when a reminder might be helpful. Suppose that my friend happens to work for another watch colony, with the queen in Seoul instead of Cupertino, for example. My friend could carry the idea back to Seoul, and within a few months, watches from a completely different colony will be reminding their wearers to breathe when they get agitated. Is this a form of procreation? Did Seoul just have sex with Cupertino? There was no direct exchange of bits, but a mutation occurred, mediated by me and my friend. Perhaps it is more like horizontal gene transfer than like sex. Horizontal gene transfer is a relatively recently discovered phenomenon where genes can migrate between species and even across domains of life, possibly mediated by viruses. More about that later.

Is it reasonable to consider my watch to be living in some sense of the word "living"? The evolutionary biologist Richard Dawkins, one of my all-time heroes, in his classic book *The Blind Watchmaker*, seems to state that it is not:

> The analogy between … watch and living organism, is false.

But here, Dawkins is referring to the fact that watches are designed by humans while living organisms evolve in a Darwinian way. He is focused only on this one aspect of "living," namely, evolution. He continues:

> All appearances to the contrary, the only watchmaker in nature is the blind forces of physics, albeit deployed in a very special way. A true watchmaker has foresight: he designs his cogs and springs, and plans their interconnections, with a future purpose in his mind's eye. Natural selection, the blind, unconscious, automatic process which Darwin discovered, and

which we now know is the explanation for the existence and apparently purposeful form of all life, has no purpose in mind. It has no mind and no mind's eye. It does not plan for the future. It has no vision, no foresight, no sight at all. If it can be said to play the role of watchmaker in nature, it is the blind watchmaker.[3]

In his zeal to debunk creationism, Dawkins seems to have, perhaps inadvertently, endowed watches with a divine creator, one with "foresight," a property that seems to lie outside the forces of physics. But aren't the humans who design watches and their foresight also forces of nature? Fortunately, later in the book, Dawkins explicitly applies evolution to technology, albeit not to watches:

Not only does the present design of a missile invite, or call forth, a suitable antidote, say a radio jamming device. The antimissile device, in its turn, invites an improvement in the design of the missile, an improvement that specifically counters the antidote, an anti-antimissile device. It is almost as though each improvement in the missile stimulates the next improvement in itself, via its effect on the antidote. Improvement in equipment feeds on itself. This is a recipe for explosive, runaway evolution.[4]

Watches are not directly trying to destroy one another, but the watch colonies headquartered in Cupertino and Seoul may be. And foresight is most certainly involved in the runaway evolutionary process of missiles and antimissile defenses.

Dawkins's point is that life was not designed by a designer that lives, somehow, *outside the system*, but rather that life was shaped by evolution and the "blind forces of physics" operating entirely *within the system*. I do not believe that Dawkins intended to state that evolution does not play a role in the design of a watch.

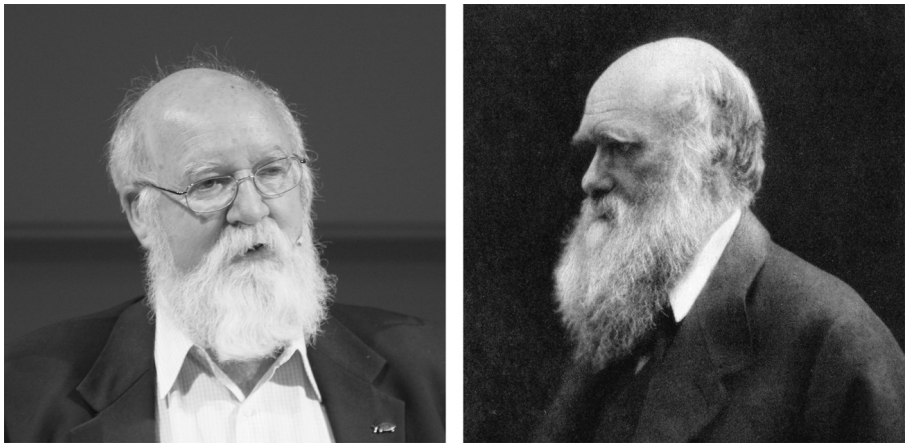
A true watchmaker is a part of nature. Unless there is something supernatural in watchmakers, they are just more complicated forces of nature. Foresight is valuable for survival and procreation, and I'm sure that

Dawkins would agree that foresight evolved in humans.[5] It was not designed. It then became a force of nature. If a watchmaker is a force of nature, then it seems reasonable to understand a watch as the result of an evolutionary process driven by forces of nature. Not even a watch has a divine creator.

A remarkable recent development in AI is that we are starting to see software designing software. Does that software have foresight? Is there something humans are capable of, when designing software, that software is not capable of? These questions are urgent and not easily answered.

BAD BOATS

Daniel Dennett, who will appear several times in this book due to his outsized influence on me, is possibly the most widely read and debated living philosopher. Working at Tufts University, the combative Dennett has taken on leading thinkers in evolutionary biology, religion, psychology, and philosophy. In what I suspect is a deliberate homage, Dennett sports a bushy beard that gives him a striking resemblance to Charles Darwin (see figure 1.4).



1.4 Daniel Dennett in 2008 and Charles Darwin in 1868. Dennett: By Mathias Schindler, CC BY-SA 3.0, via Wikimedia Commons. Darwin: By Julia Margaret Cameron, Public Domain, via Wikimedia Commons.

In his book *From Bacteria to Bach and Back*, Dennett notices that technological artifacts can exhibit a kind of procreation and mutation, following the principles of Darwinian evolution. If you will forgive my three levels of indirection, I will quote Dennett quoting Rogers and Ehrlich quoting the French philosopher known as Alain (whose real name was Émile-Auguste Chartier) writing about fishing boats in Brittany:

> Every boat is copied from another boat. … Let's reason as follows in the manner of Darwin. It is clear that a very badly made boat will end up at the bottom after one or two voyages and thus never be copied. … One could then say, with complete rigor, that it is the sea herself who fashions the boats, choosing those which function and destroying the others.[6]

A spectacular example of a badly made boat is the Swedish naval ship Vasa, which sank less than 1,500 meters into her maiden voyage from Stockholm harbor in 1628 (see figure 1.5). King Gustav II Adolf ordered her built as part of a military expansion during a war with Poland. Top heavy, with two full decks of heavy cannon and lavish adornment on a huge sterncastle, and with