FROM BACTERIA TO BACH AND BACK

THE EVOLUTION OF MINDS

Daniel C. Dennett

W. W. NORTON & COMPANY
Independent Publishers Since 1923
New York | London

Printed in the United States of America
First Edition

Manufacturing by LSC Communications, Harrisonburg, VA
Book design by Chris Welch
Production manager: Anna Oler

ISBN 978-0-393-24207-2

W. W. Norton & Company, Inc.
500 Fifth Avenue, New York, N.Y. 10110
www.wwnorton.com

W. W. Norton & Company Ltd.
15 Carlisle Street, London W1D 3BS

1 2 3 4 5 6 7 8 9 0

TO BRANDON, SAMUEL, ABIGAIL, AND ARIA

CONTENTS

Part III
TURNING OUR MINDS INSIDE OUT

LIST OF ILLUSTRATIONS

PREFACE

I started trying to think seriously about the evolution of the human mind when I was a graduate student in philosophy in Oxford in 1963 and knew almost nothing about either evolution or the human mind. In those days philosophers weren't expected to know about science, and even the most illustrious philosophers of mind were largely ignorant of work in psychology, neuroanatomy, and neurophysiology (the terms *cognitive science* and *neuroscience* would not be coined for more than a decade). The fledgling enterprise dubbed Artificial Intelligence by John McCarthy in 1956 was attracting attention, but few philosophers had ever touched a computer, whirring mysteriously in an air-conditioned prison guarded by technicians. So it was the perfect time for an utterly untrained amateur like me to get an education in all these fields. A philosopher who asked good questions about what they were doing (instead of telling them why, in principle, their projects were impossible) was apparently such a refreshing novelty that a sterling cadre of pioneering researchers took me in, gave me informal tutorials, and sent me alerts about whom to take seriously and what to read, all the while being more forgiving of my naïve misunderstandings than they would have been had I been one of their colleagues or graduate students.

Today there are dozens, hundreds, of young philosophers who do have solid interdisciplinary training in cognitive science, neuroscience, and computer science, and they are rightly held to much higher standards than I was. Some of them are my students, and even grandstudents, but other philosophers of my generation jumped into the deep end (often with more training than I) and have their own distinguished flocks of students making progress on the cutting edge, either as interdisciplinary philosophers or as philosophically trained scientists with labs of their own. They are professionals, and I am still an amateur, but by now a well-informed amateur, who gets invited to give lectures and participate in workshops and visit labs all over the world, where I continue my education, having more fun than I ever imagined an academic life could provide.

I consider this book to be, among other things, my grateful attempt to pay my tuition for all that instruction. This is what I think I've learned—a lot of it is still very conjectural, philosophical, out on a limb. I claim that it is the sketch, the backbone, of the best scientific theory to date of how our minds came into existence, how our brains work all their wonders, and, especially, how to think about minds and brains without falling into alluring philosophical traps. That is a controversial claim, of course, and I am eagerly looking forward to engaging with the reactions of both scientists and philosophers, and the amateurs who often have the most perceptive comments of all.

Many have helped me with my books, but I will concentrate here on thanking the people who have specifically helped me with the ideas in *this* book, and who, of course, are not responsible for the errors that they were unable to talk me out of. These include the participants in the Santa Fe Institute working group I organized on cultural evolution in May of 2014: Sue Blackmore, Rob Boyd, Nicolas Claidière, Joe Henrich, Olivier Morin, Pete Richerson, Peter Godfrey-Smith, Dan Sperber, and Kim Sterelny, as well as others at SFI: especially Chris Wood, Tanmoy Bhattacharya, David Wolpert, Cris Moore, Murray Gell-Mann, and David Krakauer. I would also like to express my gratitude to Louis Godbout of the Sybilla Hesse Foundation, for supporting the workshop.

Then there are my Tufts students and auditors who participated in a seminar in the spring of 2015 that went through most of the chapters in early drafts: Alicia Armijo, Edward Beuchert, David Blass, Michael Dale, Yufei Du, Brendan Fleig-Goldstein, Laura Friedman, Elyssa Harris, Justis Koon, Runeko Lovell, Robert Mathai, Jonathan Moore, Savannah Pearlman, Nikolai Renedo, Tomas Ryan, Hao Wan, Chip Williams, Oliver Yang, and Daniel Cloud, who visited the seminar to discuss his new book. And Joan Vergés-Gifra, Eric Schliesser, Pepa Toribio, and Mario Santos Sousa and the rest of the happy group who gathered at the University of Girona where I spent an intensive week as Guest Lecturer for the Ferrater Mora Chair of Contemporary Thought in May. Another test-bed was provided by Anthony Grayling and the faculty and students he has convened at the New College of the Humanities in London, where I have been trying out versions of these ideas for the last four years.

Others who wrestled with my drafts, changed my mind, noted my errors, and urged me to try for greater clarity include Sue Stafford, Murray Smith, Paul Oppenheim, Dale Peterson, Felipe de Brigard, Bryce Huebner, Enoch Lambert, Amber Ross, Justin Junge, Rosa Cao, Charles Rathkopf, Ronald Planer, Gill Shen, Dillon Bowen, and Shawn Simpson. Further good advice has come from Steve Pinker, Ray Jackendoff, David Haig, Nick Humphrey, Paul Seabright, Matt Ridley, Michael Levin, Jody Azzouni, Maarten Boudry, Krys Dolega, Frances Arnold, and John Sullivan.

As with my previous book, *Intuition Pumps and Other Tools for Thinking*, editors Drake McFeely and Brendan Curry at Norton challenged me to clarify, simplify, compress, expand, explain, and sometimes expunge, making the finished book a much more unified and effective reading experience than it would have been without their expert advice. John Brockman and Katinka Matson, as always, have been the perfect literary agents, advising, encouraging, entertaining—and, of course, selling—the author at home and abroad. Teresa Salvato, Program Coordinator at the Center for Cognitive

Studies, has handled all the logistics of my academic life for years, releasing thousands of prime-time hours for writing and researching, and played a more direct helping role for this book, tracking down books and articles in libraries and organizing the references.

Finally, my wife Susan, who has been my mainstay, advisor, critic, and best friend for more than half a century, has kept just the right amount of heat on the stove to keep the pot happily simmering, through all the ups and downs, and deserves accolades for her contributions to our joint enterprise.

Daniel Dennett
North Andover, MA
March 28, 2016

Part I

TURNING OUR WORLD UPSIDE DOWN

1

Introduction

Welcome to the jungle

How come there are minds? And how is it possible for minds to ask and answer this question? The short answer is that minds evolved and created thinking tools that eventually enabled minds to know how minds evolved, and even to know how these tools enabled them to know what minds are. What thinking tools? The simplest, on which all the others depend in various ways, are spoken words, followed by reading, writing, and arithmetic, followed by navigation and mapmaking, apprenticeship practices, and all the concrete devices for extracting and manipulating information that we have invented: compass, telescope, microscope, camera, computer, the Internet, and so on. These in turn fill our lives with technology and science, permitting us to know many things not known by any other species. We know there are bacteria; dogs don't; dolphins don't; chimpanzees don't. Even bacteria don't know there are bacteria. Our minds are different. It takes thinking tools to understand what bacteria are, and we're the only species (so far) endowed with an elaborate kit of thinking tools.

That's the short answer, and stripped down to the bare generalities it shouldn't be controversial, but lurking in the details are some surprising, even shocking, implications that aren't yet well understood or appreciated. There is a winding path leading through a jungle of science and philosophy, from the initial bland assumption

that we people are physical objects, obeying the laws of physics, to an understanding of our conscious minds. The path is strewn with difficulties, both empirical and conceptual, and there are plenty of experts who vigorously disagree on how to handle these problems. I have been struggling through these thickets and quagmires for over fifty years, and I have found a path that takes us all the way to a satisfactory—and satisfying—account of how the "magic" of our minds is accomplished without any magic, but it is neither straight nor easy. It is not the only path on offer, but it is the best, most promising to date, as I hope to show. It does require anyone who makes the trip to abandon some precious intuitions, but I think that I have at last found ways of making the act of jettisoning these "obvious truths" not just bearable but even delightful: it turns your head inside out, in a way, yielding some striking new perspectives on what is going on. But you do have to let go of some ideas that are dear to many.

There are distinguished thinkers who have disagreed with my proposals over the years, and I expect some will continue to find my new forays as outrageous as my earlier efforts, but now I'm beginning to find good company along my path, new support for my proposed landmarks, and new themes for motivating the various strange inversions of reasoning I will invite you to execute. Some of these will be familiar to those who have read my earlier work, but these ideas have been repaired, strengthened, and redesigned somewhat to do heavier lifting than heretofore. The new ones are just as counterintuitive, at first, as the old ones, and trying to appreciate them without following my convoluted path is likely to be forlorn, as I know from many years of trying, and failing, to persuade people piecemeal. Here is a warning list of some of the hazards (to comfortable thinking) you will meet on my path, and I don't expect you to "get" all of them on first encounter:

1. Darwin's strange inversion of reasoning
2. Reasons without reasoners
3. Competence without comprehension
4. Turing's strange inversion of reasoning

5. Information as design worth stealing
6. Darwinism about Darwinism
7. Feral neurons
8. Words striving to reproduce
9. The evolution of the evolution of culture
10. Hume's strange inversion of reasoning
11. Consciousness as a user-illusion
12. The age of post-intelligent design

"Information as design worth stealing? Don't you know about Shannon's mathematical theory of information?" "Feral neurons? As contrasted with what, domesticated neurons?" "Are you serious? Consciousness as an illusion? Are you kidding?"

If it weren't for the growing ranks of like-minded theorists, well-informed scientists and philosophers who agree with at least large portions of my view and have deeply contributed to it, I'd no doubt lose my nerve and decide that I was the one who's terminally confused, and of course it's possible that our bold community of enthusiasts are deluding each other, but let's find out how it goes before we chance a verdict.

I know how easy, how tempting, it is to ignore these strange ideas or dismiss them without a hearing when first encountered because I have often done so myself. They remind me of those puzzles that have a *retrospectively* obvious solution that you initially dismiss with the snap judgment: "It can't be that," or don't even consider, it is so unpromising.[1] For someone who has often accused others of mistaking failures of imagination for insights into necessity, it is embarrassing to recognize my own lapses in this regard, but having stumbled

1 One of my favorites: Four people come to a river in the night. There is a narrow bridge, and it can only hold two people at a time. They have one torch and, because it's night, the torch has to be used when crossing the bridge. Person A can cross the bridge in one minute, B in two minutes, C in five minutes, and D in eight minutes. When two people cross the bridge together, they must move at the slower person's pace. The question is, can they all get across the bridge in fifteen minutes or less?

upon (or been patiently shown) new ways of couching the issues, I am eager to pass on my newfound solutions to the big puzzles about the mind. All twelve of these ideas, and the background to render them palatable, will be presented, in *roughly* the order shown above. Roughly, because I have found that some of them defy straightforward defense: you can't appreciate them until you see what they can get you, but you can't use them until you appreciate them, so you have to start with partial expositions that sketch the idea and then circle back once you've seen it in action, to drive home the point.

The book's argument is composed of three strenuous exercises of imagination:

> turning our world upside down, following Darwin and Turing;
> then evolving evolution into intelligent design;
> and finally turning our minds inside out.

The foundation must be carefully secured, in the first five chapters, if it is to hold our imaginations in place for the second feat. The next eight chapters delve into the empirical details of the evolution of minds and language *as they appear from our inverted perspective.* This allows us to frame new questions and sketch new answers, which then sets the stage for the hardest inversion of all: seeing what consciousness looks like from the new perspective.

It's a challenging route, but there are stretches where I review familiar material to make sure everybody is on the same page. Those who know these topics better than I do can jump ahead if they wish, or they can use my treatments of them to gauge how much they should trust me on the topics they don't know much about. Let's get started.

A bird's-eye view of the journey

Life has been evolving on this planet for close to four billion years. The first two billion years (roughly) were spent optimizing the basic machinery for self-maintenance, energy acquisition and reproduction, and the only living things were *relatively* simple single-celled entities—bacteria, or their cousins, archaea: the *prokaryotes.* Then an amazing thing happened: two *different* prokaryotes, each with its own competences and habits due to its billions of years of independent evolution, collided. Collisions of this sort presumably happened countless numbers of times, but on (at least) one occasion, one cell engulfed the other, and instead of destroying the other and using the parts as fuel or building materials (eating it, in other words), it let it go on living, and, by dumb luck, found itself fitter—more competent in some ways that mattered—than it had been as an unencumbered soloist.

This was perhaps the first successful instance of *technology transfer,* a case of two different sets of competences, honed over eons of independent R&D (research and development), being united into something bigger and better. We read almost every day of Google or Amazon or General Motors gobbling up some little start-up company to get its hands on their technological innovations and savvy, advances in R&D that are easier to grow in cramped quarters than in giant corporations, but the original exploitation of this tactic gave evolution its first great boost. Random mergers don't always work out that way. In fact, they almost never work out that way, but evolution is a process that depends on amplifying things that almost never happen. For instance, mutation in DNA almost never occurs—not once in a billion copyings—but evolution depends on it. Moreover, the vast majority of mutations are either deleterious or neutral; a fortuitously "good" mutation almost never happens. But evolution depends on those rarest of rare events.

Speciation, the process in which a new species is generated when some members get isolated from their "parent" population and wander away in genetic space to form a new gene pool, is an exceedingly rare event, but the millions or billions of species that have existed on this planet each got their start with an event of speciation. Every birth in every lineage is a potential initiation of a speciation event, but speciation almost never happens, not once in a million births.

In the case we are considering, the rare improvement that resulted

from the fortuitous collision of a bacterium and an archaeon had a life-changing sequel. Being fitter, this conjoined duo reproduced more successfully than the competition, and every time it divided in two (the bacterial way of reproducing) both daughter cells included an offspring of the original guest. Henceforth their fates were joined—symbiosis—in one of the most productive episodes in the history of evolution. This was *endosymbiosis* because one of the partners was literally inside the other, unlike the side-by-side ectosymbiosis of clownfish and sea anemone or fungus and algae in lichens. Thus was born the *eukaryotic* cell, which, having more working parts, was more versatile than its ancestors, simple *prokaryotic* cells, such as bacteria.[2] Over time these eukaryotes grew much larger, more complex, more competent, *better* (the "eu" in "eukaryotic" is like the "eu" in euphonious, eulogy, and eugenics—it means *good*). Eukaryotes were the key ingredient to make possible multicellular life forms of all varieties. To a first approximation, every living thing big enough to be visible to the naked eye is a multicellular eukaryote. We are eukaryotes, and so are sharks, birds, trees, mushrooms, insects, worms, and all the other plants and animals, all direct descendants of the original eukaryotic cell.

This Eukaryotic Revolution paved the way for another great transition, the Cambrian "Explosion" more than half a billion years ago, which saw the "sudden" arrival of a bounty of new life forms. Then came what I call the MacCready Explosion, after the late great Paul MacCready, visionary engineer (and creator of the Gossamer Albatross, among other green marvels). Unlike the Cambrian diversification, which occurred over several million years about 530 million years ago (Gould 1989), the MacCready Explosion occurred in only about 10,000 years, or 500 human generations. According to MacCready's calculations (1999), at the dawn of human agriculture

2 Lane (2015) has a fascinating update (and revision) of the story of the endosymbiotic origin of eukaryotes that I have been conveying for the last twenty years or so. It now is quite secure that a bacterium and an archaeon were the Adam and Eve, not two different bacteria, as I had often said.

10,000 years ago, the worldwide human population plus their livestock and pets was only ~0.1% of the terrestrial vertebrate biomass. (We're leaving out insects, other invertebrates, and all marine animals.) Today, by his estimation, it is 98%! (Most of that is cattle.) His reflections on this amazing development are worth quoting:

> Over billions of years, on a unique sphere, chance has painted a thin covering of life—complex, improbable, wonderful and fragile. Suddenly we humans . . . have grown in population, technology, and intelligence to a position of terrible power: we now wield the paintbrush. (1999, p. 19)

There have been other *relatively* sudden changes on our planet, mass extinctions such as the Cretaceous-Paleogene extinction about sixty-six million years ago that doomed the dinosaurs, but the MacCready Explosion is certainly one of the fastest major biological changes ever to occur on Earth. It is still going on and picking up speed. We can save the planet or extinguish all life on the planet, something no other species can even imagine. It might seem obvious that the order of MacCready's three factors—population, technology, and intelligence—should be reversed: first our human *intelligence* created the *technology* (including agriculture) that then permitted the *population* boom, but as we shall see, evolution is typically an interwoven fabric of coevolutionary loops and twists: in surprising ways, our so-called native intelligence depends on both our technology and our numbers.

Our human minds are strikingly different from the minds of all other species, many times more powerful and more versatile. The long answer of how we came to have such remarkable minds is beginning to come into focus. The British biologist D'Arcy Thompson (1917) famously said, "Everything is the way it is because it got that way." Many of the puzzles (or "mysteries" or "paradoxes") of human consciousness evaporate once you ask how they could possibly have arisen—and actually try to answer the question! I mention that because some people marvel at the question and then "answer"

it by saying, "It's an impenetrable mystery!" or "God did it!" They may in the end be right, of course, but given the fabulous bounty of thinking tools recently put at our disposal and hardly used yet, this is a strikingly premature surrender. It may not be defeatist; it may be defensive. Some people would like to persuade the curious to keep their hands off the beloved mysteries, not realizing that a mystery solved is even more ravishing than the ignorant fantasies it replaces. There are some people who have looked hard at scientific explanations and disagree: to their taste, ancient myths of fiery chariots, warring gods, worlds hatching from serpent eggs, evil spells, and enchanted gardens are more delightful and worthy of attention than any rigorous, predictive scientific story. You can't please everybody.

This love of mystery is just one of the potent imagination-blockers standing in our way as we attempt to answer the question of how come there are minds, and, as I already warned, our path will have to circle back several times, returning to postponed questions that couldn't be answered until we had a background that couldn't be provided until we had the tools, which couldn't be counted on until we knew where they came from, a cycle that gradually fills in the details of a sketch that won't be convincing until we can reach a vantage point from which we can look back and see how all the parts fit together.

Douglas Hofstadter's book, *I Am a Strange Loop* (2007), describes a mind composing itself in cycles of processing that loop around, twisting and feeding on themselves, creating exuberant reactions to reflections to reminders to reevaluations that generate novel structures: ideas, fantasies, theories, and, yes, thinking tools to create still more. Read it; it will take your imagination on a roller-coaster ride, and you will learn a lot of surprising truths. My story in this book is of the larger strange looping process (composed of processes composed of processes) that generated minds like Hofstadter's (and Bach's and Darwin's) out of nothing but molecules (made of atoms made of . . .). Since the task is cyclical, we have to begin somewhere in the middle and go around several times. The

task is made difficult by a feature it doesn't share with other scientific investigations of processes (in cosmology, geology, biology, and history, for instance): people care so deeply what the answers are that they have a very hard time making themselves actually *consider* the candidate answers objectively.

For instance, some readers may already be silently shaking their heads over a claim I just made: Our human minds are strikingly different from the minds of all other species, many times more powerful and more versatile. Am I really that *prejudiced*? Am I a "species chauvinist" who actually thinks human minds are that much more wonderful than the minds of dolphins and elephants and crows and bonobos and the other clever species whose cognitive talents have recently been discovered and celebrated? Isn't this a barefaced example of the *fallacy* of "human exceptionalism"? Some readers may be ready to throw the book across the room, and others may just be unsettled by my politically incorrect lapse. It's amusing (to me, at least) that human exceptionalism provokes equal outrage in opposite directions. Some scientists and many animal lovers deplore it as an intellectual sin of the worst kind, scientifically ill-informed, an ignoble vestige of the bad old days when people routinely thought that all "dumb" animals were put on this planet for our use and amusement. Our brains are made of the same neurons as bird brains, they note, and some animal brains are just as large (and just as smart, in their own species-specific ways) as ours. The more you study the actual circumstances and behaviors of animals in the wild, the more you appreciate their brilliance. Other thinkers, particularly in the arts and humanities and social sciences, consider the *denial* of human exceptionalism to be myopic, doctrinaire, *scientism* at its worst: *Of course* our minds are orders of magnitude more powerful than the cleverest animal mind! No animal creates art, writes poetry, devises scientific theories, builds spaceships, navigates the oceans, or even tames fire. This provokes the retort: What about the elegantly decorated bowers built by bowerbirds, the political subtlety of chimpanzees, the navigational prowess of whales and elephants and migrating birds,

the virtuoso song of the nightingale, the language of the vervet monkeys, and even the honey bees? Which invites the response that these animal marvels are paltry accomplishments when compared with the genius of human artists, engineers, and scientists. Some years ago,[3] I coined the terms *romantic* and *killjoy* to refer to the sides of this intense duel over animal minds, and one of my favorite memories of this bipolar reaction to claims about animal intelligence occurred at an international scientific workshop on animal intelligence where one distinguished researcher managed to play both romantic and killjoy roles with equal passion: "Hah! You think insects are so stupid! I'll show you how smart they are. Consider this result. . . . !" Followed later on the same day by, "So you think bees are so clever? Let me show you how stupid they *really* are! They're mindless little robots!"

Peace! We will see that both sides are right about some things and wrong about others. We're not the Godlike geniuses we sometimes think we are, but animals are not so smart either, and yet both humans and (other) animals are admirably equipped to deal "brilliantly" with many of the challenges thrown at them by a difficult, if not always cruel, world. And our human minds are uniquely powerful in ways that we can begin to understand once we see how they got that way.

Why do we care so much? That is one of the many hanging questions that needs an answer, but not right now, except in briefest outline: While the processes that gave rise to this caring go back thousands of years, and in some regards millions or even billions of years, they first became a *topic*—an object to think about and care

3 Since this book is the culmination of a half century of work on these topics, normal academic citation practices would sprinkle its pages with dozens of interruptions of the "(Dennett 1971, 1991, 2013)" variety, but such voluminous self-citation would send the wrong message. My thinking has been shaped by hundreds of thinkers, and I have tried to credit the key sources on all the ideas discussed as they arise, while sequestering most of the information about where I myself have expanded on these points to the Appendix: The Background (p. 415), for the convenience of anyone who is curious to see how the arguments developed.

about—at the birth of modern science in the seventeenth century, so that is where I will break into the ring and start this version of the story.

The Cartesian wound

Si, abbiamo un anima. Ma é fatta di tanti piccoli robot!
(Yes, we have a soul, but it's made of lots of tiny robots!)
—Headline for an interview with me by Giulio Giorello
in *Corriere della Serra*, Milan, 1997

René Descartes, the seventeenth-century French scientist and philosopher, was very impressed with his own mind, for good reason. He called it his *res cogitans*, or thinking thing, and it struck him, on reflection, as a thing of miraculous competence. If anybody had the right to be in awe of his own mind, Descartes did. He was undoubtedly one of the greatest scientists of all time, with major work in mathematics, optics, physics, and physiology; and the inventor of one of the most valuable thinking tools of all time, the system of "Cartesian coordinates" that enables us to translate between algebra and geometry, paving the way for calculus and letting us plot almost anything we want to investigate, from aardvark growth to zinc futures. Descartes propounded the original TOE (theory of everything), a prototypical Grand Unified Theory, which he published under the immodest title *Le Monde* (*The World*). It purported to explain everything from the orbits of the planets and the nature of light to the tides, from volcanoes to magnets, why water forms into spherical drops, how fire is struck from flint, and much, much more. His theory was almost all dead wrong, but it held together surprisingly well and is strangely plausible even in today's hindsight. It took Sir Isaac Newton to come up with a better physics, in his famous *Principia*, an explicit refutation of Descartes's theory.

Descartes didn't think it was just his mind that was wonderful; he thought that all normal human minds were wonderful, capable

of feats that no mere animal could match, feats that were beyond the reach of any imaginable *mechanism*, however elaborate and complicated. So he concluded that minds like his (and yours) were not material entities, like lungs or brains, but made of some *second* kind of stuff that didn't have to obey the laws of physics—articulating the view known as *dualism*, and, often, *Cartesian dualism*. This idea that mind isn't matter and matter can't be mind was not invented by Descartes. It had seemed obvious to reflective people for thousands of years that our minds are not like the furniture of the "external" world. The doctrine that *each of us has an immaterial (and immortal) soul that resides in and controls the material body* long passed for shared knowledge, thanks to the instruction of the Church. But it was Descartes who distilled this default assumption into a positive "theory": The immaterial mind, the conscious *thinking thing* that we know intimately through introspection, is somehow in communication with the material brain, which provides all the input *but none of the understanding or experience.*

The problem with dualism, ever since Descartes, is that nobody has ever been able to offer a convincing account of how these postulated interactive transactions between mind and body could occur without violating the laws of physics. The candidates on display today offer us a choice between a revolution in science so radical that it can't be described (which is convenient, since critics are standing by, ready to pounce) or a declaration that some things are just Mysteries, beyond human understanding (which is also convenient if you don't have any ideas and want to exit swiftly). But even if, as I noted years ago, dualism tends to be regarded as a cliff over which you push your opponents, those left on the plateau have a lot of unfinished business constructing a theory that is *not* dualism in disguise. The mysterious linkage between "mind and matter" has been a battleground of scientists and philosophers since the seventeenth century.

Francis Crick, the recently deceased co-discoverer of the structure of DNA, was another of history's greatest scientists, and his last major piece of writing was *The Astonishing Hypothesis: The Scientific Search for the Soul* (1994), in which he argued that dualism is false; the mind just *is* the brain, a material organ with no mysterious extra properties not found in other living organisms. He was by no means the first to put forward this denial of dualism; it has been the prevailing—but not unanimous—opinion of both scientists and philosophers for the better part of a century. In fact, many of us in the field objected to his title. There was nothing astonishing about this hypothesis; it had been our working assumption for decades! Its *denial* would be astonishing, like being told that gold was not composed of atoms or that the law of gravity didn't hold on Mars. Why should anyone expect that *consciousness* would bifurcate the universe dramatically, when even *life* and *reproduction* could be accounted for in physico-chemical terms? But Crick wasn't writing his book for scientists and philosophers, and he knew that among laypeople, the appeal of dualism was still quite overpowering. It seemed not only obvious to them that their private thoughts and experiences were somehow conducted in some medium *in addition to* the neuronal spike trains scientists had found buzzing around in their brains, but the prospect of denying dualism threatened horrible consequences as well: If "we are just machines," what happens to free will and responsibility? How could our lives have meaning at all if we are just huge collections of proteins and other molecules churning away according to the laws of chemistry and physics? If moral precepts were nothing but extrusions generated by the hordes of microbiological nano-machines between our ears, how could they make a difference worth honoring?

Crick did his best to make "the astonishing hypothesis" not just comprehensible but also palatable to the lay public. Despite his clear and energetic writing, and unparalleled gravitas, he didn't make much progress. This was largely, I think, because in spite of his book's alarm bell of a title, he underestimated the emotional turmoil this idea provokes. Crick was an excellent explainer of science to nonscientists, but the pedagogical problems in this arena are not the usual ones of attracting and holding the attention of semi-bewildered and intimidated laypeople and getting them to

work through a smattering of math. When the topic of consciousness arises, the difficult task is to keep a lid on the anxieties and suspicions that seduce people—including many scientists—into distorting what we know and aiming preemptive strikes at dangerous ideas they dimly see looming. Moreover, on this topic *everybody's an expert.* People are calmly prepared to be instructed about the chemical properties of calcium or the microbiological details of cancer, but they think they have a particular personal authority about the nature of their own conscious experiences that can trump any hypothesis they find unacceptable.

Crick is not alone. Many others have tried their hand at knitting up what one of the best of them, Terrence Deacon, has called "the Cartesian wound that severed mind from body at the birth of modern science" (2011, p. 544). Their efforts are often fascinating, informative, and persuasive, but no one has yet managed to be entirely *convincing.* I have devoted half a century, my entire academic life, to the project in a dozen books and hundreds of articles tackling various pieces of the puzzle, without managing to move all that many readers from wary agnosticism to calm conviction. Undaunted, I am trying once again and going for the whole story this time.

Why do I think it is worth trying? Because, first, I think we have made tremendous scientific progress in the last twenty years; many of the impressionistic hunches of yore can now be replaced with well-researched details. I plan to rely heavily on the bounty of experimental and theoretical work that others have recently provided. And second, I think I now have a better sense of the various undercurrents of resistance that shackle our imaginations, and I plan to expose and disarm them as we go, so that, for the first time, the doubters can *take seriously* the prospect of a scientific, materialist theory of their own minds.

Cartesian gravity

Over the years, trudging back and forth over the battleground, participating in many skirmishes, I've gradually come to be able to

see that there are powerful forces at work, distorting imagination—my own imagination included—pulling us first one way and then another. If you learn to see these forces too, you will find that suddenly things begin falling into place in a new way. You can identify the forces tugging at your thinking, and then set up alarms to alert you and buffers to protect you, so that you can resist them effectively while simultaneously exploiting them, because they are not just distorting; they can also be imagination-enhancing, launching your thinking into new orbits.

One cold, starry night over thirty years ago, I stood with some of my Tufts students looking up at the sky while my friend, the philosopher of science, Paul Churchland instructed us how to *see the plane of the ecliptic,* that is, to look at the other visible planets in the sky and picture them, and ourselves, as wheeling around the sun all on the same invisible plane. It helps to tip your head just so and remind yourself of where the sun must be, way back behind you. Suddenly, the orientation clicks into place and shazam, you *see* it![4] Of course we all knew for years that this was the situation of our planet in the solar system, but until Paul made us see it, it was a rather inert piece of knowledge. Inspired by his example, I am going to present some eye-opening (actually *mind*-opening) experiences that I hope will move your mind into some new and delightful places.

The original distorting force, which I will call *Cartesian gravity,* actually gives birth to several other forces, to which I will expose you again and again, in different guises, until you can see them clearly too. Their most readily "visible" manifestations are already familiar to most everyone—too familiar, in fact, since we tend to think we have already taken their measure. We underestimate them. We must look behind them, and beyond them, to see the way they tend to sculpt our thinking.

Let's begin by looking back at Crick's "astonishing hypothesis." Those of us who insist that we don't find it at all astonishing fuel

4 Churchland includes instruction and a diagram that will help you enjoy this delightful effect in his 1979 book *Scientific Realism and the Plasticity of Mind.*

our confidence by reminding ourselves of the majestic array of well-solved puzzles, well-sleuthed discoveries, well-confirmed theories of modern, materialistic science that we all take for granted these days. When you think about it, it is just amazing how much we human beings have figured out in the few centuries since Descartes. We know how atoms are structured, how chemical elements interact, how plants and animals propagate, how microscopic pathogens thrive and spread, how continents drift, how hurricanes are born, and much, much more. We know our brains are made of the same ingredients as all the other things we've explained, and we know that we belong to an evolved lineage that can be traced back to the dawn of life. If we can explain *self-repair in bacteria* and *respiration in tadpoles* and *digestion in elephants*, why shouldn't *conscious thinking in H. sapiens* eventually divulge its secret workings to the same ever-improving, self-enhancing scientific juggernaut?

That's a rhetorical question, and trying to *answer* rhetorical questions instead of being cowed by them is a good habit to cultivate. So might consciousness be more challenging than self-repair or respiration or digestion, and if so, why? Perhaps because it *seems* so different, so private, so intimately *available* to each of us in a way unlike any other phenomenon in our living bodies. It is not all that hard these days to imagine how respiration works even if you're ignorant of the details: you breathe in the air, which we know is a combination of different gases, and we breathe out what we can't use—carbon dioxide, as most people know. One way or another the lungs must filter out and grab what is needed (oxygen) and exude the waste product (carbon dioxide). Not hard to grasp in outline. The phenomenon of smelling a cookie and suddenly remembering an event in your childhood seems, in contrast, not at all mechanical. "Make me a nostalgia-machine!" "What? What could the parts possibly do?" Even the most doctrinaire materialists will admit that they have only foggy and programmatic ideas about how brain activities might amount to nostalgia or wistfulness or prurient curiosity, for example.

Not so much an astonishing hypothesis, many might admit, as a

dumbfounding hypothesis, a hypothesis about which one can only wave one's hands and hope. Still, it's a comfortable position to maintain, and it's tempting to diagnose those who disagree—the self-appointed Defenders of Consciousness from Science—as suffering from one or another ignominious failing: narcissism ("I refuse to have *my* glorious mind captured in the snares of science!"); fear ("If my mind is just my brain, I won't be in charge; life will have no meaning!"); or disdain ("These simple-minded, *scientistic* reductionists! They have no idea how far short they fall in their puny attempts to appreciate the world of meaning!").

These diagnoses are often warranted. There is no shortage of pathetic bleats issuing from the mouths of the Defenders, but the concerns that motivate them are not idle fantasies. Those who find Crick's hypothesis not just astonishing but also deeply repugnant are onto something important, and there is also no shortage of anti-dualist philosophers and scientists who are not yet comfortable with materialism and are casting about for something in between, something that can actually make some progress on the science of consciousness without falling into either. The trouble is that they tend to misdescribe it, inflating it into something deep and metaphysical.[5]

What they are feeling is a way of thinking, an overlearned habit, so well entrenched in our psychology that denying it or abandoning it is literally unthinkable. One sign of this is that the confident scientific attitude expressed by the "other side" begins to tremble the closer the scientists get to a certain set of issues dealing with consciousness, and they soon find themselves, in spite of themselves, adopting the shunned perspective of the Defenders. I am going to

5 Working with Nick Humphrey some years ago on what was then called multiple personality disorder, I discovered the almost irresistible temptation, even in Nick and myself, to exaggerate anything that strikes one as both important and uncanny. I surmise that whenever we human beings encounter something truly strange and unsettling, our attempts to describe *to ourselves* what we are experiencing *tend* to err on the side of exaggeration, perhaps out of a subliminal desire to impress ourselves with the imperative that this is something we ignore at our peril and must get to the bottom of.

describe this dynamic process metaphorically at the outset to provide a simple framework for building a less metaphorical, more explicit and factual understanding of what is happening.

Suppose the would-be mind-explainer starts with her *own* mind. She stands at Home, on Planet Descartes, meditating on the task ahead and looking at the external universe from the "first-person point of view." From this vantage point, she relies on all the familiar furniture of her mind to keep her bearings, and Cartesian gravity is the force that locks her into this egocentric point of view "from the inside." Her soliloquy might be, echoing Descartes: "Here I am, a conscious thinking thing, intimately acquainted with the ideas in my own mind, which I know better than anybody else just because they're mine." She cannot help but be a Defender of her own Home. Meanwhile, from faraway comes the scientific explorer of consciousness, approaching Planet Descartes confidently, armed with instruments, maps, models, and theories, and starts moving in for the triumphant conquest. The closer she gets, however, the more uncomfortable she finds herself; she is being dragged into an orientation she knows she must avoid, but the force is too strong. As she lands on Planet Descartes she finds herself flipped suddenly into first-person orientation, feet on the ground but now somehow unable to reach, or use, the tools she brought along to finish the job. Cartesian gravity is all but irresistible when you get that close to the surface of Planet Descartes. How did she get there, and what happened in that confusing last-minute *inversion*? (Strange inversions will be a major theme in this book.) There seem to be two competing orientations, the first-person point of view of the Defenders and the third-person point of view of the scientists, much like the two ways of seeing the philosophers' favorite illusions, the duck-rabbit and the Necker cube. You can't adopt both orientations at once.

The problem posed by Cartesian gravity is sometimes called the Explanatory Gap (Levine 1983) but the discussions under that name strike me as largely fruitless because the participants tend to see it as a chasm, not a glitch in their imaginations. They may have *discovered* the "gap," but they don't see it for what it actually is

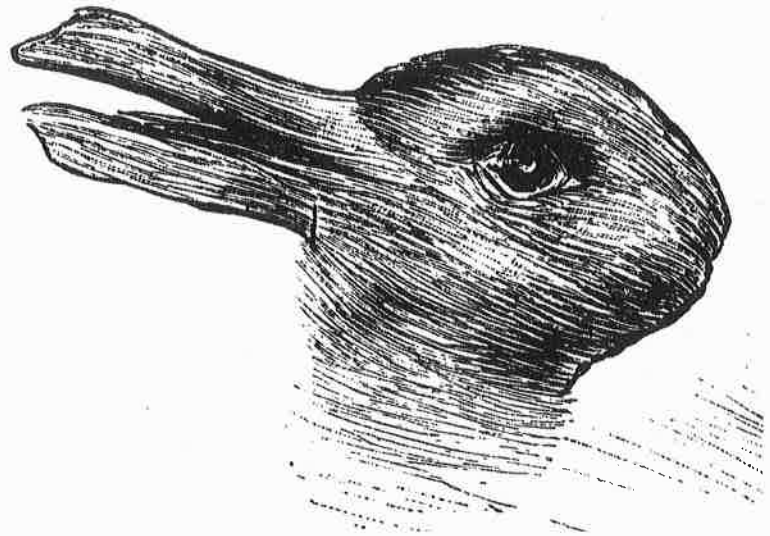


FIGURE 1.1: Duck-rabbit.

because they haven't asked "how it got that way." By reconceiving of the gap as a dynamic imagination-distorter that has arisen for good reasons, we can learn how to traverse it safely or—what may amount to the same thing—make it vanish.

Cartesian gravity, unlike the gravity of physics, does not act on things in proportion to their mass and proximity to other massy items; it acts on ideas or representations of things in proportion to their proximity *in content* to other ideas that play privileged roles in the maintenance of a living thing. (What this means will gradually become clear, I hope—and then we can set this metaphorical way of speaking aside, as a ladder we have climbed and no longer need to rely on.) The *idea* of Cartesian gravity, as so far presented, is just a metaphor, but the phenomenon I am calling by this metaphorical name is perfectly real, a disruptive force that bedevils (and sometimes aids) our imaginations, and unlike the gravity of physics, it is itself an evolved phenomenon. In order to understand it, we need to ask how and why it arose on planet Earth.

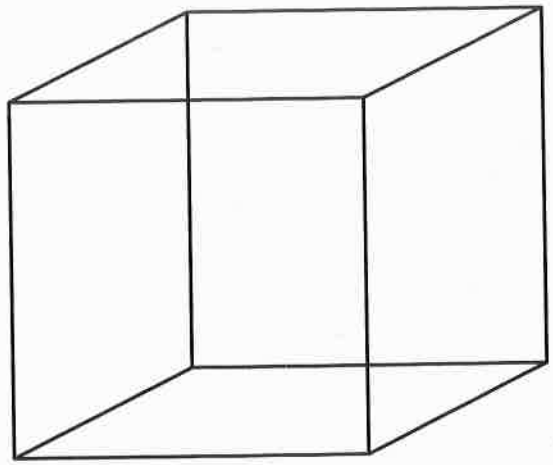


FIGURE 1.2: Necker cube.

It will take several passes over the same history, with different details highlighted each time, to answer this question. We tend to underestimate the strength of the forces that distort our imaginations, especially when confronted by irreconcilable insights that are "undeniable." It is not that we *can't* deny them; it is that we *won't* deny them, won't even *try* to deny them. Practicing on the forces that are easy to identify—species chauvinism, human exceptionalism, sexism—prepares us to recognize more subtle forces at work. In the next chapter I turn briefly to the very beginning of life on the planet and give a preliminary outline of the story to come, with scant detail, and also tackle one of the first objections that (I predict) will occur to the reader while encountering that outline. I cast evolutionary processes as *design* processes (processes of research and development, or R&D) and this *adaptationist* or *reverse-engineering* perspective has long lived under an undeserved cloud of suspicion. Contrary to widespread belief, as we shall see, adaptationism is alive and well in evolutionary biology.

2

Before Bacteria and Bach

Why Bach?

To get a sound perspective on our history, we actually have to go back to the time before bacteria, before life in any form existed, since some of the demands on getting life going at all send important echoes down the eons to explain features of our minds today. And before turning to *that* story, let me pause to draw attention to one word: "Bach." I could have chosen "From Archaea to Shakespeare" or "From *E. coli* to Einstein" or perhaps "From Prokaryotes to Picasso," but the alliteration of "Bacteria to Bach" proved irresistible.

What about the glaring fact that all the candidates I just considered in my pantheon of great minds are men? What an awkward stumble out of the starting blocks! Do I really want to alienate many readers at the outset? What was I thinking? I did it on purpose, to provide a relatively mild and simple example of *one variety* of the gravitational forces of Cartesian gravity we will deal with. If you bristled when you noticed my all-male roster of genius, then good. That means you won't forget that I've taken out a loan on your patience, to be repaid at some later time in the book. Bristling (like any acute emotional reaction, from fear to amusement) is a sort of boldface for your memory, making the offending item less likely to be forgotten. At this point I ask you to resist the urge

to launch a preemptive strike. Throughout our journey, we will need to identify uncomfortable facts without indulging in premature explanations or rebuttals. While I delight in having readers who are not only paying close attention but also way ahead of me, I would much prefer that you bide your time, cutting me some slack—giving me enough rope to hang myself, if you like—instead of trying to derail my attempt at a calm, objective account with your premonitions.

So let's stand down, pause, and review a few plain facts, leaving the explanations and refutations for later. It is an obvious fact that although there have been many brilliant women of great attainment, none of them has achieved the iconic status of Aristotle, Bach, Copernicus, Dickens, Einstein. . . . I could easily list a dozen more men in the same league, but try for yourself to think of a great female thinker who could readily displace any of these men in playing the emblematic role in my title. (My favorites would be Jane Austen, Marie Curie, Ada Lovelace, and Hypatia of Alexandria. I doubt I've overlooked any obvious candidates, but time will tell.)

There have *not yet* been any female superstar geniuses. What might explain this fact? Political oppression? The self-fulfilling sexist prophecies that rob young girls of inspiring role models? Media bias over the centuries? Genes? Please don't jump to conclusions, even if you think the answer is obvious. (I don't.) We will soon see that genes, though essential players in the history of mind, are nowhere near as important as many like to think. Genes may account for basic animal competence, but *genes don't account for genius!* Moreover, the traditional view of great human societies owing their greatness to the creative brilliance of (some of) their inhabitants is, I will try to show, just about backward. Human culture *itself* is a more fecund generator of brilliant innovations than any troupe of geniuses, of either gender. This it achieves by a process of *cultural* evolution that is as much "the author" of our finest achievements as any individual thinker is.

The very idea that evolution by natural selection might have a foundational role to play in understanding human culture fills some people, even wise and thoughtful people, with loathing. They see human culture as something transcendent, the miraculous endowment that distinguishes us human beings from the beasts, the last bulwark against creeping reductionism, against genetic determinism, against the philistinism they think they see in contemporary science. And, in reaction to these "culture vultures," there are hard-nosed scientists to whom any appeal to "culture" smells of mystery-mongering or worse.

When I hear the word "culture" I reach for my gun.[6]

Now I must ask "both sides" to holster their guns and be patient. There is a middle ground that can do justice to the humanities and science at the same time, explaining how human culture got up and running by a process of evolution of cultural items—*memes*—that invaded human brains the way viruses invade human bodies. Yes, memes have *not* been shown to be a bad idea, and they will get their day in court in this book. Those on both sides who scoff and hoot at the idea of memes will see that there are *good* objections to memes—in addition to the flawed "refutations" that have been uncritically accepted by many who just can't stand the idea—but these good objections point to refinements that save the concept of memes after all.

So whose side am I on? Readers intent on framing the issue in these terms have missed the point. This polarization of visions, with cheering, hissing spectators egging on the combatants, is just a conveniently obvious first manifestation of the forces I am trying to render visible to all and neutralize. There is much more to come—subtler and more insidious pressures on the thinking of scientists, philosophers, and laypeople alike. Now back to my first pass at the story.

6 *Not* Hermann Göring (and not Heinrich Himmler either). According to Wikipedia, this oft-misattributed declaration was born as a line of dialogue in a pro-Nazi play by Hans Johst.

How investigating the prebiotic world is like playing chess

The simplest, earliest life form capable of reproducing itself, something like a bacterium, was already a breathtakingly complex and brilliantly designed self-maintaining system. (Hang on. Did I not just give aid and comfort to the Intelligent Design crowd? No. But how can a good materialist, atheist Darwinian like me keep a straight face while declaring the earliest reproducing life forms to be *brilliantly designed*? Hold your fire.)

A well-known chicken-and-egg-problem beloved by the Intelligent Design folks challenges us to undo the "paradox" of the origin of life: evolution by natural selection couldn't get started until there were *reproducing* things, since there would be no offspring to inherit the best designs, but the simplest reproducing thing is much too complex to arise by mere chance.[7] So, it is claimed, evolution cannot get started without a helping hand from an Intelligent Designer. This is a defective argument, a combination of misdirection and failure of imagination, as we shall see. But we should concede the truly striking fact that the first concoction of molecules capable of reliably reproducing itself was—must have been—an "engineering" marvel, composed of thousands of complex parts working together.

For researchers working on the origin of life this poses a straightforward challenge: How could this *possibly* come about without a miracle? (Perhaps an intelligent designer from another galaxy was responsible, but this would just postpone the question and make it harder to address.) The way to proceed is clear: start with the minimal specification for a living, reproducing thing—a list of all the things it *has to be able to do*—and work backward, making an inventory of the available raw materials (often called the feedstock molecules of prebiotic chemistry), and asking what sequence of

7 I will have a lot to say about intelligent design—by human beings—in this book but almost nothing more about Intelligent Design, the latest wave of creationist propaganda. It is not worth any further rebuttal.

possible events could gradually bring together, non-miraculously, all the necessary parts in the right positions to accomplish the job. Notice that this minimal specification lists what the thing must do, a list of functions, not a list of parts and materials. A mousetrap has to trap mice, and a can-opener has to open cans, and a living thing has to capture energy and protect (or repair) itself long enough to reproduce.

How could such a living thing possibly arise? If you can answer this question, you *win*, rather like achieving checkmate in chess. This is a huge undertaking, with many gaps still to fill, but every year there are encouraging breakthroughs, so confidence runs high that the task can be done, the game can be won. There may actually be many ways life could possibly have arisen out of nonlife, but finding just one that deserves scientific allegiance (until a better alternative is discovered) would muffle the "impossible in principle" choir for good. However, finding even one way is a task so daunting that it has fueled the conviction among researchers that even though the processes that must be invoked to create the end product are utterly blind and purposeless, the product itself is not just intricate but stunningly effective at what it does—a brilliant design. It takes all the ingenuity human reverse engineers can muster to figure out how the thing got assembled. A commentary by Jack Szostak, one of the leading researchers on one of the biggest breakthroughs of recent years (by Powner, Gerland, and Sutherland 2009), illustrates the attitude perfectly. (Don't worry about the chemical details mentioned; just see how the investigation is conducted, as revealed by the phrases I have italicized.)

> For 40 years, efforts to understand the prebiotic synthesis of the ribonucleotide *building blocks* of RNA have been *based on the assumption that they must have assembled from their three molecular components*: a nucleobase (which can be adenine, guanine, cytosine or uracil), a ribose sugar and phosphate. Of the many difficulties encountered by those in the field, the most frustrating has been *the failure to find any way of properly joining the*

> *pyrimidine nucleobases—cytosine and uracil—to ribose.* . . . But Powner et al. revive the prospects of the "RNA first" model by exploring a pathway for pyrimidine ribonucleotide synthesis in which the sugar and nucleobase emerge from a common precursor. In this pathway, the complete ribonucleotide structure forms without using free sugar and nucleobase molecules as intermediates. This central insight, combined with a series of additional innovations, *provides a remarkably efficient solution to the problem of prebiotic ribonucleotide synthesis.* (Szostak 2009)

Greg Mayer (2009), an evolutionary biologist commenting on this, makes an important point:

> John Sutherland, one of Powner's coauthors, and in whose lab the work was done, worked on the problem for twelve years before he found the solution. What if he had given up after ten? Could we have concluded that no synthesis was possible? No. This work demonstrates the futility of all the various sorts of arguments—the argument from design, the God of the gaps, the argument from personal incredulity—that rely on ignorance as their chief premise.

Throughout this book I will exploit the perspective of *reverse engineering*, taking on the premise that every living thing is a product of nonmysterious physical processes that gradually brought all the elements together, refining them along the way, and eventually arrived at the working system we observe, or at some hypothesized intermediate system, a stepping-stone that would represent clear *progress* toward the living things we know exist. The cascade of processes must make changes that we can see, in retrospect, to be *improvements* in the design of the emerging systems. (We're on the way to checkmate. Are we making progress?) Until there were systems that could be strictly called *reproducing systems*, the processes at work were only proto-evolutionary, semi-Darwinian, partial *analogues* of proper evolution by natural selection; they were processes

that raised the likelihood that various combinations of ingredients would arise and persist, concentrating the feedstock molecules until this eventually led to the origin of life. A living thing must capture *enough* energy and materials, and fend off its own destruction *long enough* to construct a *good enough* replica of itself. The reverse-engineering perspective is ubiquitous in biology and is obligatory in investigations of the origin of life. It always involves some kind of optimality considerations: What is the *simplest* chemical structure that could *possibly* do *x*? Or would phenomenon *x* be *stable enough* to sustain process *y*?

In a highly influential essay, Stephen Jay Gould and Richard Lewontin (1979) coined the phrase "Panglossian paradigm" as a deliberately abusive term for the brand of biology—adaptationism—that relies on the methodological principle of assuming, until proven otherwise, that all the parts of an organism are *good for something*. That is, they have useful roles to play, such as pumping blood, improving speed of locomotion, fending off infection, digesting food, dissipating heat, attracting mates, and so forth. The assumption is built right into the reverse-engineering perspective that sees all living things as efficiently composed of parts with functions. (There are well-recognized exceptions: for instance, features that *used* to be good for something and are now vestigial, along for the ride unless they are too expensive to maintain, and features that have no actual function but just "drifted" to "fixation" by chance.)

Gould and Lewontin's joke was a recycled caricature. In *Candide*, Voltaire created Dr. Pangloss, a wickedly funny caricature of the philosopher Leibniz, who had maintained that our world was the *best* of all possible worlds. In Dr. Pangloss's overfertile imagination, there was no quirk, no deformity, no catastrophe of Nature that couldn't be seen, in retrospect, to have a function, to be a blessing, just what a benevolent God would arrange for us, the lucky inhabitants of the perfect world. Venereal disease, for instance, "is indispensable in this best of worlds. For if Columbus, when visiting the West Indies, had not caught this disease, which poisons the source of generation, which frequently even hinders generation, and is clearly opposed to

the great end of Nature, we should have neither chocolate nor cochineal" (quoted by Gould and Lewontin 1979, p. 151). Leibniz scholars will insist, with some justice, that Voltaire's parody is hugely unfair to Leibniz, but leave that aside. Was Gould and Lewontin's reuse of the idea an unfair caricature of the use of optimality assumptions in biology? Yes, and it has had two unfortunate effects: their attack on adaptationism has been misinterpreted by some evolution-dreaders as tantamount to a refutation of the theory of natural selection, and it has convinced many biologists that they should censor not only their language but also their thinking, as if reverse engineering was some sort of illicit trick they should shun if at all possible.

Those working on the origin of life have ignored the "Pangloss" critique of their methods, knowing that their strategic assumptions serve to direct the investigation away from fruitless wandering. There is no point in looking at chemical reactions that couldn't possibly generate a target structure presumed to be a necessary component. Admittedly, there are risks to this strategy; as Szostak notes, for years the researchers made the mistaken assumption that the obviously best, most efficient way of uniting the nucleobases to the ribose was directly, and they overlooked the more devious path of having the ribonucleotide emerge from a common precursor, without involving the intermediate steps that had *seemed* necessary.

In chess, a *gambit* is a strategy that gives up material—a step backward, it seems—in order to take a better, forward step from an improved position. When trying to calculate what your opponent is going to do, gambits are hard to spot, since they seem at first to be losing moves that can be safely ignored, since one's opponent is not that stupid. The same risk of ignoring devious but fruitful trails besets the reverse engineer in biology, since, as Francis Crick famously said, enunciating what he called Orgel's Second Rule: "Evolution is cleverer than you are." The uncanny way the blind, purposeless churn of evolution (including prebiotic chemical evolution) uncovers off-the-wall solutions to problems is not evidence for an Intelligent Designer, nor is it grounds for *abandoning* reverse

engineering, which would mean giving up the inquiry altogether; it is grounds for persisting and improving your reverse-engineering game. As in chess, don't give up; learn from your mistakes and keep on exploring, as imaginatively as you can, bearing in mind that your hypotheses, however plausible, still risk disconfirmation, which should be conscientiously sought.

Here is an example of a possible *gambit* in the origin of life. It is initially tempting to assume that the very first living thing capable of reproducing must have been the *simplest possible* living thing (given the existing conditions on the planet at the time). First things first: Make the simplest replicator you can imagine and then build on that foundation. But this is by no means necessary. It is possible, and more likely, I think, that a rather inelegantly complicated, expensive, slow, Rube-Goldberg conglomeration of *objets trouvés* was the first real replicator, and after it got the replication ball rolling, this ungainly replicator was repeatedly simplified in competition with its kin. Many of the most baffling magic tricks depend on the audience not imagining the ridiculously extravagant lengths magicians will go to in order to achieve a baffling effect. If you want to reverse engineer magicians, you should always remind yourself that they have no shame, no abhorrence of bizarre expenditures for "tiny" effects that they can then exploit. Nature, similarly, has no shame—and no budget, and all the time in the world.

Talk of improvements or progress in the slow, uncertain process of biogenesis is not indulging in illicit value judgments (which have no place in science, let's agree) but is rather an acknowledgment of the ever-present requirements of stability and efficiency in anything living. If you like, you can imagine biochemists working on how something utterly *terrible* might come into existence, a doomsday device or self-replicating death ray. They would still have to discipline their search by imagining possible paths to construct this horror. And they might well marvel at the brilliance of the design they finally figured out. I will have more to say about the presuppositions and implications of reverse engineering in biology later. Here I hope to forestall premature dismissal of my project by any who has

been persuaded, directly or by hearsay, that Gould and Lewontin's propaganda against adaptationism was fatal. Contrary to the opinion widely engendered by their famous essay, adaptationism is alive and well; reverse engineering, when conducted with due attention to the risks and obligations, is still the royal road to discovery in biology and the only path to discovery in the demanding world of prebiotic chemistry of the origin of life.[8]

Next I want to look at the phenomenon of the origin of life from a more philosophical perspective, as the origin of *reasons*. Is there design in Nature or only apparent design? If we consider evolutionary biology to be a species of reverse engineering, does this imply that there are reasons for the arrangements of the parts of living things? Whose reasons? Or can there be reasons without a reasoner, designs without a designer?

8 Nikolai Renedo has suggested to me that the take-home message of Gould and Lewontin's famous essay is "be on the lookout for gambits," which is certainly good advice for any adaptationist to follow. If that is what Gould and Lewontin intended, however, they failed to convey it to their audiences, both lay and scientific, where the opinion persists that it was an authoritative demotion of adaptationism as a central feature of evolutionary thinking.

3

On the Origin of Reasons

The death or rebirth of teleology?

Darwin is often credited with overthrowing Aristotle's all-too-influential doctrine that everything in the world has a purpose, an "end" (in the sense that the ends justify the means), or as the French say, a *raison d'être*, a reason for being.

Aristotle identified four questions we might want to ask about anything:

1. What is it made of, or its *material cause*?
2. What is its structure, or its *formal cause*?
3. How did it get started, or what is its *efficient cause*?
4. What is its purpose, or its *final*, or *telic*, *cause*?

The Greek for the fourth cause is *telos* from which we derive the term *teleology*. Science, we are often told, has banished the *telos*, and we have Darwin to thank for this. As Karl Marx (1861) once famously put it, in his appreciation of Darwin's *Origin of Species*: "Not only is a death blow dealt here for the first time to 'Teleology' in the natural sciences but their rational meaning is empirically explained."

But a closer look shows that Marx is equivocating between two views that continue to be defended:

> We should banish all teleological formulations from the natural sciences
>
> *or*
>
> now that we can "empirically explain" the "rational meaning" of natural phenomena without ancient ideology (of entelechies, Intelligent Creators and the like), we can replace old-fashioned teleology with new, post-Darwinian teleology.

This equivocation is firmly knitted into the practice and declarations of many thoughtful scientists to this day. On the one hand, biologists routinely and ubiquitously refer to the *functions* of behaviors such as foraging and territory marking, organs such as eyes and swim bladders, subcellular "machinery" such as ribosomes, chemical cycles such as the Krebs cycle, and macromolecules such as motor proteins and hemoglobin. But some thoughtful biologists and philosophers of biology are uneasy about these claims and insist that all this talk of functions and purposes is really only shorthand, a handy metaphor, and that strictly speaking there are no such things as functions, no purposes, no teleology at all in the world. Here we see the effect of another of the imagination-distorting forces spawned by Cartesian gravity. So seductive is the lure of Cartesian thinking that in order to resist it, some think we should follow the abstemious principle that whenever there is any risk of *infection* by prescientific concepts—of souls and spirits, Aristotelian teleology and the like—it is best to err on the side of squeaky-clean, absolute quarantine. This is often a fine principle: the surgeon excising a tumor takes out a generous "margin" around the suspect tissue; political leaders institute buffer zones, to keep dangerous weapons—or dangerous ideologies—at arm's length.

A little propaganda can help keep people vigilant. Among the epithets hurled at unrepentant teleologists are "Darwinian paranoia" (Francis 2004; Godfrey-Smith 2009) and "conspiracy theorists" (Rosenberg 2011). It is of course open to defend an intermediate

position that forbids certain teleological excesses but licenses more staid and circumscribed varieties of talk about functions, and philosophers have devised a variety of such views. My informal sense is that many scientists assume that some such sane middle position is in place and must have been adequately defended in some book or article that they probably read years ago. So far as I know, however, no such consensus classic text exists,[9] and many of the scientists who guiltlessly allude to the functions of whatever they are studying still insist that they would *never* commit the sin of teleology.

One of the further forces in operation here is the desire not to give aid and comfort to creationists and the Intelligent Design crowd. By speaking of purpose and design in Nature, we (apparently) give them half their case; it is better, some think, to maintain a stern embargo on such themes and insist that *strictly speaking* nothing in the biosphere is designed unless it is designed by human artificers. Nature's way of generating complex systems (organs, behaviors, etc.) is so unlike an artificer's way that we should not use the same language to describe them. Thus Richard Dawkins speaks (on occasion—e.g., 1996, p. 4) of *designoid* features of organisms, and in *The Ancestors' Tale* (2004) he says, "the illusion of design conjured by Darwinian natural selection is so breathtakingly powerful" (p. 457). I disagree with this overkill austerity, which can backfire badly. A few years ago I overheard some Harvard Medical School students in a bar marveling at the intricacies to be found in the protein machinery inside a cell. One of them exclaimed, "How could anybody believe in evolution in the face of all that design!" The others did not demur, whatever their private thoughts. Why would anyone say that? Evolutionists aren't embarrassed by the intricacy of Nature. They revel in it! Discovering and explaining the evolution

9 Biologists and philosophers have written often about function talk, and although there are persistent disagreements about how to license it, there is something of a consensus that evolutionary considerations do the trick one way or another for natural functions, and facts about both history and current competence anchor attributions of function to artifacts. For a good anthology of the best work by both biologists and philosophers, see Allen, Bekoff, and Lauder 1998.

of the intracellular complexities that govern the life of a cell has been one of the glories of evolutionary microbiology in recent years. But this fellow's remark suggests that one of the themes gaining ground in common understanding is that evolutionary biologists are reluctant to "admit" or "acknowledge" the manifest design in Nature. People should know better, especially medical students!

Consider in this regard Christoph Schönborn, Catholic archbishop of Vienna, who was seduced by the Intelligent Design folks into denying the theory of natural selection on the grounds that it couldn't explain all the design. He said, notoriously, in a *New York Times* op-ed piece entitled "Finding Design in Nature" (July 7, 2005):

> The Catholic Church, while leaving to science many details about the history of life on earth, proclaims that by the light of reason the human intellect can readily and clearly discern purpose and design in the natural world, including the world of living things. Evolution in the sense of common ancestry might be true, but evolution in the neo-Darwinian sense—an unguided, unplanned process of random variation and natural selection—is not. Any system of thought that denies or seeks to explain away the overwhelming evidence for design in biology is ideology, not science.

Which battle do we want to fight? Do we want to try to convince lay people that they don't really see the design that is stunningly obvious at every scale in biology, or would we rather try to persuade them that what Darwin has shown is that there can be design—real design, as real as it gets—without an Intelligent Designer? We have persuaded the world that atoms are not atomic, and that the Earth goes around the Sun. Why shrink from the pedagogical task of showing that there can be design without a designer? So I am defending here (once again, with new emphasis) the following claim:

> *The biosphere is utterly saturated with design, with purpose, with reasons.* What I call the "design stance" predicts and explains features throughout the living world using the same assumptions that work so well when reverse-engineering artifacts made by (somewhat) intelligent human designers.

There are three different but closely related strategies or stances we can adopt when trying to understand, explain, and predict phenomena: the physical stance, the design stance, and the intentional stance (Dennett 1971, 1981, 1983, 1987, and elsewhere). The physical stance is the least risky but also the most difficult; you treat the phenomenon in question as a physical phenomenon, obeying the laws of physics, and use your hard-won understanding of physics to predict what will happen next. The design stance works only for things that are designed, either artifacts or living things or their parts, and have functions or purposes. The intentional stance works *primarily* for things that are designed to use information to accomplish their functions. It works by treating the thing as a rational agent, attributing "beliefs" and "desires" and "rationality" to the thing, and predicting that it will act rationally.

Evolution by natural selection is not itself a designed thing, an agent with purposes, but it acts as if it were (it occupies the role vacated by the Intelligent Designer): it is a set of processes that "find" and "track" reasons for things to be arranged one way rather than another. The chief difference between the reasons found by evolution and the reasons found by human designers is that the latter are typically (but not always) represented in the minds of the designers, whereas the reasons uncovered by natural selection are represented for the first time by those human investigators who succeed in reverse engineering Nature's productions. Dawkins's title, *The Blind Watchmaker* (1986), nicely evokes the apparently paradoxical nature of these processes: on the one hand they are blind, mindless, without goals, and on the other hand they produce designed entities galore, many of which become competent artificers (nest-builders, web-spinners, and so forth) and a few become intelligent designers and builders: us.

Evolutionary processes brought purposes and reasons into exis-

tence the same way they brought color vision (and hence colors) into existence: gradually. If we understand the way our human world of reasons grew out of a simpler world where there were no reasons, we will see that purposes and reasons are as real as colors, as real as life. Thinkers who insist that Darwin has banished teleology should add, for consistency's sake, that science has also demonstrated the unreality of colors and of life itself. Atoms are all there is, and atoms aren't colored, and aren't alive either. How could mere large conglomerations of uncolored, unalive things add up to colored, live things? This is a rhetorical question that should be, and can be, answered (eventually). Now I want to defend the claim that there are reasons for what proteins do, and there are reasons for what bacteria do, what trees do, what animals do, and what we do. (And there are colors as well, of course, and yes, Virginia, life really exists.)

Different senses of "why"

Perhaps the best way of seeing the reality, indeed the ubiquity in Nature, of *reasons* is to reflect on the different meanings of "why." The English word is equivocal, and the main ambiguity is marked by a familiar pair of substitute phrases: *what for?* and *how come?*"

> "Why are you handing me your camera?" asks *what* are you doing this *for*?
>
> "Why does ice float?" asks *how come*: what it is about the way ice forms that makes it lower density than liquid water?

The *how come* question asks for a *process narrative* that explains the phenomenon without saying it is *for* anything. "Why is the sky blue?" "Why is the sand on the beach sorted by size?" "Why did the ground just shake?" "Why does hail accompany thunderstorms?" "Why is this dry mud cracked in such a fashion?" And also, "Why did this turbine blade fail?" Some folks might wish to treat the question of why ice floats as inviting a *what for* reason—God's reason, presumably—for this feature of the inanimate world. ("I guess God wanted fish to be able to live under the ice in the winter, and if ponds froze from the bottom up, this would be hard on the fish.") But as long as we have an answer to the *how come* question, in terms of physics and chemistry, it really would be something like paranoia to ask for more.

Compare four questions:

1. Do you know the reason why planets are spherical?
2. Do you know the reason why ball bearings are spherical?
3. Do you know the reason why asteroids aren't spherical?
4. Do you know the reason why dice aren't spherical?

The word "reason" is acceptable in all four questions (at least to my ear—how about yours?), but the answers to (1) and (3) don't give *reasons* (there aren't any reasons); they give *causes*, or process narratives. In some contexts the word "reason" can mean *cause*, unfortunately. You can answer questions (2) and (4) with process narratives along the lines of "well, the ball bearings were made on a lathe of sorts, which spun the metal . . . and the dice were cast in boxlike molds . . ." but those are not *reasons*. Sometimes people confuse the different questions, as in a memorable exchange that occurred in a debate I had with an ardent champion of Skinnerian behaviorism, Lou Michaels, at Western Michigan University in 1974. I had presented my paper "Skinner Skinned" (in *Brainstorms* 1978), and Michaels, in his rebuttal, delivered a particularly bold bit of behaviorist ideology, to which I responded, "But why do you say that, Lou?" to which his instant reply was "Because I have been rewarded for saying that in the past." I was demanding a reason—a *what for*—and getting a process narrative—a *how come*—in reply. There is a difference, and the Skinnerians' failed attempt to make it go away should alert positivistically minded scientists that they pay a big price in understanding if they try to banish "what for."

The first two sentences of this book are "How come there are

minds? And how is it possible for minds to ask and answer this question?" It is asking for a process narrative, and that is what I am going to provide. But it will be a process narrative that also answers the questions how come there are "what for?" questions, and what are "what for?" questions for?

The evolution of "why": from *how come* to *what for*

Evolution by natural selection starts with *how come* and arrives at *what for*. We start with a lifeless world in which there are no reasons, no purposes at all, but there are processes that happen: rotating planets, tides, freezing, thawing, volcanic eruptions, and kazillions of chemical reactions. Some of those processes happen to generate other processes that happen to generate other processes until at some "point" (but don't look for a bright line) we *find it appropriate* to describe the *reasons* why some things are arranged as they now are. (*Why* do we find it appropriate, and how did we get into that state of mind? Patience, the answer to that will come soon.)

A central feature of human interaction, and one of the features unique to our species, is the activity of asking others to explain themselves, to justify their choices and actions, and then judging, endorsing, rebutting their answers, in recursive rounds of the "why?" game. Children catch on early, and often overdo their roles, trying the patience of their parents. "Why are you sawing the board?" "I'm making a new door." "Why are you making a new door?" "So we can close the house up when we go out." "Why do you want to close the house up when we go out?" . . . "Why don't we want strangers taking our things?" . . . "Why do we have things?" The fluency with which we all engage in this mutual reason-checking testifies to its importance in conducting our lives: our capacity to *respond* appropriately in this reason-checking activity is the root of *responsibility*. (Anscombe 1957) Those who cannot explain themselves or cannot be moved by the reasons offered by others, those who are "deaf to" the persuasions of advisors, are rightly judged to be of diminished responsibility and are treated differently by the law.

This activity of demanding and evaluating each other's reasons for action does not occupy our every waking hour, but it does play a major role in coordinating our activities, initiating the young into their adult roles, and establishing the norms by which we judge one another. So central is this practice to our own way of life that it is sometimes hard to imagine how other social species—dolphins, wolves, and chimpanzees, for instance—can get along without it. How do the juveniles "learn their place," for instance, without being *told* their place? How do elephants settle disagreements about when to move on or where to go next? Subtle instinctual signals of approval and disapproval must suffice, and we should also remember that no other species engages in the level of complex cooperative behaviors that we human beings have achieved.

Wilfrid Sellars, a philosopher at the University of Pittsburgh, described this activity of reasoning with one another as creating or constituting "the logical space of reasons" (1962) and inspired a generation of Pittsburgh philosophers, led by Robert Brandom and John Haugeland, to explore this arena in detail. What are the permissible moves, and why? How do new considerations enter the space, and how are transgressions dealt with? The space of reasons is bound by *norms*, by mutual recognition of how things *ought* to go—the right way, not the wrong way, to play the reason-giving game. Wherever there are reasons, then, there is room for, and a need for, some kind of *justification* and the possibility of *correction* when something goes wrong.

This "normativity" is the foundation of ethics: the ability to appreciate how reason-giving *ought to go* is a prerequisite for appreciating how life in society ought to go. Why and how did this practice and its rules arise? It hasn't existed forever, but it exists now. How come and what for? The Pittsburgh philosophers have not addressed this question, asking how "it got that way," so we will have to supplement their analysis with some careful speculation of our own on the evolution of the reason-giving game. I will try to show

that ignoring this question has led the Pittsburgh philosophers to elide the distinction between two different kinds of norms and their associated modes of correction, which I will call *social normativity* and *instrumental normativity*. The former, analyzed and celebrated at Pittsburgh, is concerned with the *social* norms that arise within the practice of communication and collaboration (hence Haugeland [1998] speaks of the "censoriousness" of members of society as the force that does the correcting). The latter is concerned with quality control or efficiency, the norms of engineering, you could say, as revealed by market forces or by natural failures. This is nicely illustrated by the distinction between a good deed and a good tool. A good deed might be clumsily executed and even fail in its purpose, while a good tool might be an efficient torture device or evil weapon. We can see the same contrast in negative cases, in the distinction between *naughty* and *stupid*. People may punish you for being naughty, by their lights, but Nature itself may mindlessly punish you for being stupid. As we shall see, we need both kinds of norms to create the perspective from which reasons are *discernible* in Nature.

Reason-appreciation did *not* coevolve with reasons the way color vision coevolved with color. Reason-appreciation is a later, more advanced product of evolution than reasons.

Wherever there are reasons, an implicit norm may be invoked: real reasons are supposed always to be good reasons, reasons that justify the feature in question. (No demand for justification is implied by any "how come" question.) When we reverse engineer a newly discovered artifact, for instance, we may ask why there is a conspicuous knob in the corner that doesn't seem to "do anything" (anything useful—it makes a shadow when light falls on it, and changes the center of gravity of the artifact, but has no apparent function). We expect, until we learn otherwise, that the designer had a reason, a good reason, for that knob. It might be that there used to be a good reason, but that reason has lapsed and the manufacturers have forgotten this fact. The knob is vestigial, functionless, and present only because of inertia in the manufacturing process. The same expectations drive the reverse-engineering explorations of living things, and biologists often permit themselves to speak, casually, about what "Nature intended" or what "evolution had in mind" when it "selected" some puzzling feature of a living thing.[10] No doubt the biologists' practice is a direct descendant of the reverse engineering of artifacts designed and made by other human beings, which is itself a direct descendant of the societal institution of asking for and giving reasons for human activities. That *might* mean that this practice is an outdated vestige of prescientific thinking—and many biologists surmise as much—or it might mean that biologists have found a brilliant extension of reverse engineering into the living realm, using the thinking tools Nature has endowed us with to discover real patterns in the world that can well be called the *reasons* for the existence of other real patterns. To defend the latter claim, we need to take a look at how evolution itself could get going.

Go forth and multiply

In *Darwin's Dangerous Idea* (1995), I argued that natural selection is an *algorithmic* process, a collection of sorting algorithms that are themselves *composed* of generate-and-test algorithms that exploit randomness (pseudo-randomness, chaos) in the generation phase, and some sort of mindless quality-control testing phase, with the winners advancing in the tournament by having more offspring. How does this cascade of generative processes get under way? As noted in the last chapter, the actual suite of processes that led to the origin of life are still unknown, but we can dissipate some of the fog

10 For instance, biologist Shirley Tilghman, in the 2003 Watson Lecture, said: "But clearly, what is immediately apparent when you look at any part of those two genomes that have been compared is that evolution has indeed been hard at work, conserving far more of the genome than we could explain by genes and their closely allied regulatory elements. . . . Scientists should have a field day trying to understand what evolution had in mind when she paid so much attention to these little segments of DNA."

by noting that, as usual, a variety of gradual processes of revision are available to get the ball rolling.

The prebiotic or abiotic world was not utter chaos, a random confetti of atoms in motion. In particular there were *cycles*, at many spatio-temporal scales: seasons, night and day, tides, the water cycle, and thousands of chemical cycles discoverable at the atomic and molecular level. Think of cycles as "do-loops" in algorithms, actions that return to a starting point after "accomplishing" something—accumulating something, or moving something, or sorting something, for instance—and then repeating (and repeating and repeating), gradually changing the conditions in the world and thus *raising the probability that something new will happen.* A striking abiotic example is illustrated by Kessler and Werner in *Science* 2003.

These stone circles would strike anyone as a highly improbable scattering of stones across the landscape; it looks "man-made"—reminiscent of the elegant outdoor sculptures by Andy Goldsworthy—but it is the natural outcome of hundreds or thousands of mindless cycles of freezing and thawing on Spitsbergen in the Arctic. New England farmers have known for centuries about frost driving a "fresh crop" of stones up to the soil surface every winter; stones that have to be removed before plowing and planting. The classic New England "stone walls" we still see today along field edges and marching through former fields now reforested, were never meant to keep anything in or out; they are really not walls but very long narrow piles of boulders and small rocks hauled to the nearest part of the edge of the cultivated field. They are clearly the result of deliberate, hard human work, which had a purpose. Ironically, if the farmers hadn't removed the stones, over many cycles of freezing and thawing the stones might have formed one of the "patterned ground" phenomena illustrated here, not always circles, but more often polygons, and sometimes labyrinths and other patterns. Kessler and Werner provide an explanation of the process with a model—an algorithm—that produces these different sorting effects by varying the parameters of stone size, soil moisture and

FIGURE 3.1: Kessler and Werner, stone circles. © *Science magazine and Mark A. Kessler.*

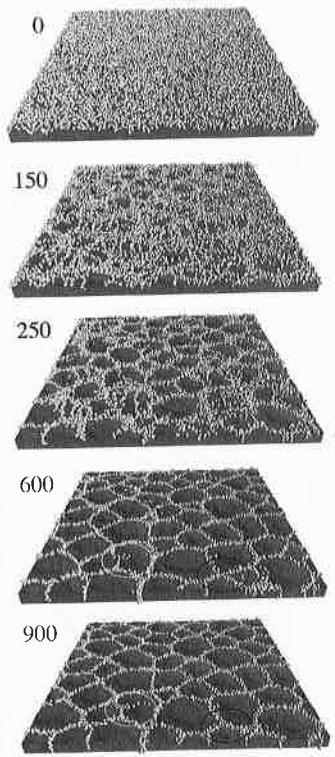


FIGURE 3.2: Kessler and Werner's stone-sorting algorithm at work. © *Science magazine and Mark A. Kessler.*

density, temperature, speed of freezing, and hillslope gradient. So we have a pretty good idea *how come* these phenomena exist where they do, and anybody who encountered these stone circles and concluded that there had to be a purposeful artificer behind them, an answer to the *what for* question, would be wrong.

In the abiotic world, many similar cycles occur concurrently but asynchronously, wheels within wheels within wheels, with different periods of repetition, "exploring" the space of chemical possibility. This would be a variety of parallel processing, a little bit like the *mass production* of industry, which makes lots of different parts in different places at different rates of production and then brings them together for assembly, except that this abiotic mass production is utterly unplanned and unmotivated.

There is no differential *re-production* in the abiotic world, but we do get varieties of differential *persistence*: some temporary combinations of parts hang around longer than others, thereby having more time to pick up revisions and adjustments. The rich can get richer, in short, even though they can't yet bequeath their riches to descendants. Differential persistence must then somehow gradually turn into differential *reproduction*. The proto-Darwinian algorithms of differential "survival" of chemical combinations can give rise to auto-catalytic reaction cycles that in turn give rise to differential *replication* as just a special case of differential *persistence*, a very special case, a particularly explosive type that multiplies its advantage by . . . multiplication! It generates many near-duplicate persisters, which can then "explore" many more slightly different corners of the world than any one or two persisters could do.

"A diamond is forever" according to the advertising slogan, but that is an exaggeration. A diamond is magnificently persistent, much more persistent than its typical competition, but its persistence is simply modeled by its linear descent through time, Tuesday's diamond being like its parent, Monday's diamond, and so forth. It never multiplies. But it can accumulate changes, wear and tear, a coating of mud that hardens, and so forth, which may make it more or less persistent. Like other durable things, it is affected by many cycles, many do-loops that involve it in one way or another. Usually these effects do not accumulate for long but rather get wiped out by later effects, but sometimes a barrier happens to get erected: a wall or membrane of some sort that provides extra shielding.

In the world of software, two well-recognized phenomena are *serendipity* and its opposite *clobbering*. The former is the chance collision of two unrelated processes with a happy result, and clobbering is such a collision with a destructive result. Walls or membranes that tend for whatever reason to prevent clobbering will be particularly persistent and will permit internal cycles (do-loops) to operate without interference. And so we see the engineering necessity of membranes to house the collection of

chemical cycles—the Krebs cycle and *thousands* of others—that together permit life to emerge. (An excellent source on this algorithmic view of chemical cycles in cells is Dennis Bray, *Wetware* 2009.) Even the simplest bacterial cells have a sort of nervous system composed of chemical networks of exquisite efficiency and elegance. But how could just the right combination of membranes and do-loops ever arise in the prebiotic world? "Not in a million years!" some say. Fair enough, but then how about once in a hundred million years? It only has to happen once to ignite the fuse of reproduction.

Imagine we are back in the early days of this process where persistence turns gradually into multiplication, and we see a proliferation of some type of items where before there were none and we ask, "Why are we seeing these improbable things here?" The question is equivocal! For now there is both a process narrative answer, *how come,* and a justification, *what for.* We are confronting a situation in which some chemical structures are present while chemically possible alternatives are absent, and what we are looking at are things that are *better* at persisting in the local circumstances than their contemporary alternatives. *Before we can have competent reproducers, we have to have competent persisters, structures with enough stability to hang around long enough to pick up revisions.* This is not a very impressive competence, to be sure, but it is just what the Darwinian account needs: something that is only sorta competent, nothing too fancy. We are witnessing an "automatic" (algorithmic) paring away of the *nonfunctional,* crowded out by the functional. And by the time we get to a reproducing bacterium, there is functional virtuosity galore. In other words, there are *reasons why* the parts are shaped and ordered as they are. We can reverse engineer any reproducing entity, determining its good and its bad, and saying *why* it is good or bad. This is the birth of reasons, and it is satisfying to note that this is a case of what Glenn Adelson has aptly called Darwinism about Darwinism (Godfrey-Smith 2009): we see the gradual emergence of the species of reasons out of the species of mere causes, *what fors* out of *how comes,* with no "essential" dividing line between them. Just as there is no Prime Mammal—the first mammal that didn't have a mammal for a mother—there is no Prime Reason, the first feature of the biosphere that helped something exist because it made it better at existing than the "competition."

Natural selection is thus an automatic reason-finder, which "discovers" and "endorses" and "focuses" reasons over many generations. The scare quotes are to remind us that natural selection doesn't have a mind, doesn't itself have reasons, but is nevertheless competent to perform this "task" of design refinement. Let's be sure we know how to cash out the scare quotes. Consider a population of beetles with lots of variation in it. Some do well (at multiplying); most do not. Take the minority (typically) that do well, reproductively, and ask about each one: *why* did it do better than average. Our question is equivocal; it can be interpreted as asking *how come* or *what for.* In many cases, most cases, the answer is *no reason at all;* it's just dumb luck, good or bad. In which case we can have only a *how come* answer to our question. But if there is a subset, perhaps very small, of cases in which there is an answer to the *what for* question, a *difference that happens to make a difference,* then those cases have in common the germ of a reason, a proto-reason, if you like. The process narrative explains how it came about and also, in the process, points to why these are better than those, why they won the competition. "Let the best entity win!" is the slogan of the evolution tournament, and the winners, being better, wear the justification of their enhancements on their sleeves. In every generation, in every lineage, only some competitors manage to reproduce, and each descendant in the next generation is either just lucky or lucky-to-be-gifted in some way. The latter group was *selected* (for cause, you might say, but better would be for a *reason*). This process accounts for the accumulation of function by a process that blindly tracks reasons, creating things that have purposes but don't need to know them. The Need to Know principle made famous in spy novels also reigns in the biosphere: an organ-

ism doesn't need to know the reasons why the gifts it inherits are beneficial to it, and natural selection itself doesn't need to know what it's doing.

Darwin understood this:

> The term "natural selection" is in some respects a bad one, as it seems to imply conscious choice; but this will be disregarded after a little familiarity. No one objects to chemists speaking of "elective affinity"; and certainly an acid has no more choice in combining with a base, than the conditions of life have in determining whether or not a new form be selected or preserved. . . . For brevity sake I sometimes speak of natural selection as an intelligent power;—in the same way as astronomers speak of the attraction of gravity as ruling the movements of the planets. . . . I have, also, often personified Nature; for I have found it difficult to avoid this ambiguity; but I mean by nature only the aggregate action and product of many natural laws,—and by laws only the ascertained sequence of events. (1868, pp. 6–7)

So there were reasons long before there were reason-representers—us. The reasons tracked by evolution I have called "free-floating rationales," a term that has apparently jangled the nerves of some few thinkers, who suspect I am conjuring up ghosts of some sort. Not at all. Free-floating rationales are no more ghostly or problematic than numbers or centers of gravity. Cubes had eight corners before people invented ways of articulating arithmetic, and asteroids had centers of gravity before there were physicists to dream up the idea and calculate with it. Reasons existed long before there were reasoners. Some find this way of thinking unnerving and probably "unsound," but I am not relenting. Instead I am hoping here to calm their fears and convince them that we should all be happy to speak of the reasons uncovered by evolution before they were ever expressed or represented by human investigators or

any other minds.[11] Consider the strikingly similar constructions in figures 3.3 and 3.4 of the color insert following p. 238.

The termite castle and Gaudí's La Sagrada Familia are very similar in shape but utterly different in genesis and construction. *There are reasons* for the structures and shapes of the termite castle, but they are not represented by any of the termites who constructed it. There is no Architect Termite who planned the structure, nor do any individual termites have the slightest clue about why they build the way they do. This is competence without comprehension, about which more later. There are also reasons for the structures and shapes of Gaudí's masterpiece, but they are (in the main) Gaudí's reasons. Gaudí *had* reasons for the shapes he ordered created; *there are* reasons for the shapes created by the termites, but the termites didn't *have* those reasons. There are reasons why trees spread their branches, but they are not in any strong sense the trees' reasons. Sponges do things for reasons, bacteria do things for reasons; even viruses do things for reasons. But they don't *have* the reasons; they don't need to have the reasons.

Are *we* the only reason-representers? This is a very important question, but I will postpone an answer until I have provided a wider setting for the perspective shift I am proposing here. So far, what I take to have shown is that Darwin didn't extinguish teleology; he naturalized it, but this verdict is not as widely accepted as it should be, and a vague squeamishness leads some scientists to go overboard avoiding design talk and reason talk. The space of reasons is created by the human practice of reason-giving and is bound by norms, both social/ethical and instrumental (the difference between being naughty and being stupid). Reverse engineering in biology is a descendant of reason-giving-judging.

11 Philosophers who are skeptical about my intransigence on this score might like to read T. M. Scanlon's recent book, *Being Realistic about Reasons* (2014), for an exhaustive and exhausting survey of the problems one confronts if one ignores engineering reasons and concentrates on *having* moral reasons for action.

The evolution of *what for* from *how come* can be seen in the way we interpret the gradual emergence of living things via a cascade of prebiotic cycles. Free-floating rationales emerge as the reasons why some features exist; they do not presuppose intelligent designers, even though the designs that emerge are extraordinarily good. For instance, there are reasons why termite colonies have the features they do, but the termites, unlike Gaudí, do not have or represent reasons, and their excellent designs are not products of an intelligent designer.

4

Two Strange Inversions of Reasoning

How Darwin and Turing broke a spell

The world before Darwin was held together not by science but by tradition. All things in the universe, from the most exalted ("man") to the most humble (the ant, the pebble, the raindrop) were the creations of a still more exalted thing, God, an omnipotent and omniscient intelligent creator—who bore a striking resemblance to the second-most exalted thing. Call this the trickle-down theory of creation. Darwin replaced it with the bubble-up theory of creation. Robert MacKenzie Beverley,[12] one of Darwin's nineteenth-century critics, put it vividly:

> In the theory with which we have to deal, Absolute Ignorance is the artificer; so that we may enunciate as the fundamental principle of the whole system, that, IN ORDER TO MAKE A PERFECT AND BEAUTIFUL MACHINE, IT IS NOT REQUISITE TO KNOW HOW TO MAKE IT. This proposition will be found, on careful examination, to express, in condensed form, the essential purport of the Theory, and to express in a few

12 I have been misidentifying this author as Robert Beverley MacKenzie for over thirty years; I thank the fact checkers at Norton for correcting my error.

> words all Mr. Darwin's meaning; who, by a strange inversion of reasoning, seems to think Absolute Ignorance fully qualified to take the place of Absolute Wisdom in all the achievements of creative skill. (Beverley 1868)

This was indeed a "strange inversion of reasoning" and the incredulity expressed by Beverley is still echoing through a discouragingly large proportion of the population in the twenty-first century.

When we turn to Darwin's bubble-up theory of creation, we can conceive of all the creative design work metaphorically as lifting in what I call Design Space. It has to start with the first crude replicators, as we saw in chapter 3, and gradually ratchet up, by wave after wave of natural selection, to multicellular life in all its forms. Is such a process really capable of having produced all the wonders we observe in the biosphere? Skeptics ever since Darwin have tried to demonstrate that one marvel or another is simply unapproachable by this laborious and unintelligent route. They have been searching for something alive but *unevolvable*. My term for such a phenomenon is a *skyhook*, named after the mythical convenience you can hang in the sky to hold up your pulley or whatever you want to lift (Dennett 1995). A skyhook floats high in Design Space, unsupported by ancestors, the direct result of a special act of intelligent creation. And time and again, these skeptics have discovered not a miraculous skyhook but a wonderful *crane*, a nonmiraculous innovation in Design Space that enables ever more efficient exploration of the possibilities of design, ever more powerful lifting in Design Space. Endosymbiosis is a crane; it lifted simple single cells into a realm of much complexity, where multicellular life could take off. Sex is a crane; it permitted gene pools to be stirred up, and thus much more effectively sampled by the blind trial-and-error processes of natural selection. Language and culture are cranes, evolved novelties that opened up vast spaces of possibility to be explored by ever more intelligent (but not miraculously intelligent) designers. Without the addition of language and culture to the arsenal of R&D tools available to evolution, there wouldn't be glow-in-the-dark tobacco plants with firefly genes in them. These are not

miraculous. They are just as clearly fruits of the Tree of Life as spider webs and beaver dams, but the probability of their emerging without the helping hand of *Homo sapiens* and our cultural tools is nil.

As we learn more and more about the nano-machinery of life that makes all this possible, we can appreciate a second strange inversion of reasoning, achieved almost a century later by another brilliant Englishman: Alan Turing. Here is Turing's strange inversion, put in language borrowed from Beverley:

> IN ORDER TO BE A PERFECT AND BEAUTIFUL COMPUTING MACHINE, IT IS NOT REQUISITE TO KNOW WHAT ARITHMETIC IS.

Before Turing's invention there were computers, by the hundreds or thousands, employed to work on scientific and engineering calculations. Computers were people, not machines. Many of them were women, and many had degrees in mathematics. They were human beings who knew what arithmetic was, but Turing had a great insight: they didn't need to know this! As he noted, "The behavior of the computer at any moment is determined by the symbols which he is observing, and his 'state of mind' at that moment" (Turing 1936, 5). That "state of mind" (in Turing's scare quotes) was a dead-simple set of if-then instructions about what to do and what "state of mind" to go into next (and repeat until you see the instruction to STOP). Turing showed that it was possible to design mindless machines that were Absolutely Ignorant, but that could do arithmetic perfectly, following "instructions" that could be mechanically implemented. More importantly, he showed that if their instructions included *conditional branching* (if-then instructions, such as "if you observe 0, replace it with 1 and move left, and if you observe 1 leave it as is and move right, and change to state *n*."), then these machines could pursue indefinitely complex paths determined by the instructions, which gave them a remarkable competence: they could do *anything* computational. In other words, a programmable digital computer is a Universal Turing Machine, capable of mimicking any

special-purpose digital computer by following a set of instructions that implement that special-purpose computer in software.[13] (You don't have to rewire your smartphone to get it to do new tasks; just download an app and turn it into a star finder or translator or hand calculator or spell-checker or. . . .) A huge Design Space of information-processing was made accessible by Turing, and he foresaw that there was a traversable path from Absolute Ignorance to Artificial Intelligence, a long series of lifting steps in that Design Space.

Many people can't abide Darwin's strange inversion. We call them creationists. They are still looking for skyhooks—"irreducibly complex" (Behe 1996) features of the biosphere that could not have evolved by Darwinian processes. Many more people can't abide Turing's strange inversion either, and for strikingly similar reasons. They want to believe that the wonders of the mind are inaccessible by mere material processes, that minds are, if not literally miraculous, then mysterious in ways that defy natural science. They don't want the Cartesian wound to be healed.

Why not? We've already noted some of their less presentable motives: fear, pride, the misplaced love of unsolved mystery. Here is another reason (is it *how come* or *what for*?): Both Darwin and Turing claim to have discovered something truly unsettling to a human mind—*competence without comprehension*. Beverley expressed his outrage with gusto: the *very idea* of creative skill without intelligence! Consider how shockingly this goes against an idea enshrined in our educational policies and practices: *comprehension as the (best) source of competence*. We send our children to universities so that they will gain an understanding of all the ways the world works that will stand them in good stead throughout their lives, generating competences

13 The standard jargon for asserting this is known as the Church-Turing Thesis, formulated by logician Alonzo Church: "all effective procedures are Turing-computable"—though of course many of them are not feasible since they take too long to run. Since our understanding of what counts as an effective procedure (basically, a computer program or algorithm) is unavoidably intuitive, this thesis cannot be proved, but it is almost universally accepted, so much so that Turing-computability is typically taken as an acceptable operational definition of effectiveness.

as needed from the valuable store of comprehension we have inculcated in them. (I am using "comprehension" and "understanding" as synonymous, by the way, favoring "comprehension" for its alliteration in the slogan, which will come up again and again.) Why do we disparage rote learning these days? Because we have seen—haven't we?—that getting children to *understand* a topic or a method is the way (the only way or just the best way?) to make them competent with regard to that topic or method. We disparage the witless memorizer who just fills in the blanks on the template without knowing what the point is. We scoff at the idea that paint-by-numbers kits are the way to train creative artists. Our motto might well be

If you make them comprehend, their competence will follow!

Note that there is more than a smidgen of ideology at play here. We are quite familiar with some disastrous misapplications of our hallowed principle, such as the "new math," which tried—unsuccessfully—to teach children set theory and other abstract concepts first, instead of drilling them on addition and subtraction, the multiplication table, fractions, and simple algorithms like long division, or counting by twos and fives and tens.

The armed forces are some of the most effective educational institutions in the world, turning average high school students into reliable jet-engine mechanics, radar operators, navigators, and a host of other technical specialists thanks to heavy doses of "drill and practice." In due course, a valuable variety of comprehension arises out of the instilled competences in these practiced practitioners, so we have good empirical evidence that competence doesn't always depend on comprehension and sometimes is a precondition for comprehension. What Darwin and Turing did was envisage the most extreme version of this point: *all* the brilliance and comprehension in the world arises ultimately out of uncomprehending competences compounded over time into ever more competent—and *hence* comprehending—systems. This is indeed a strange inversion, overthrowing the pre-Darwinian mind-first vision of Creation

with a mind-*last* vision of the eventual evolution of us, intelligent designers at long last.

Our skepticism about competence without comprehension has causes, not reasons. It doesn't "stand to reason" that there cannot be competence without comprehension; it just feels right, and it feels right *because* our minds have been shaped to think that way. It took Darwin to break the spell cast by that way of thinking, and Turing shortly thereafter came along and broke it again, opening up the novel idea that we might invert the traditional order and build comprehension out of a cascade of competences in much the way evolution by natural selection builds ever more brilliant internal arrangements, organs, and instincts without having to comprehend what it is doing.

There is one big difference between Darwin's strange inversion and Turing's. Darwin showed how brilliant designs could be created by cascades of processes lacking all intelligence, but the system for Turing's cascades of processes was the product of a very intelligent designer, Turing. One might say that while Darwin *discovered* evolution by natural selection, Turing *invented* the computer. Many people contend that an intelligent God had to set up all the conditions for evolution by natural selection to occur, and Turing appears to be playing that role in setting up the underlying idea of a (material, non-living, non-comprehending) computer which can then become the arena in which comprehension might arise by something a little bit like evolution, a series of design improvements concocted from the basic building blocks of computation. Doesn't Turing's role as intelligent designer *oppose* rather than *extend* the reach of Darwin's strange inversion? No, and answering this important question is a major task for the rest of the book. The short explanation is that Turing himself is one of the twigs on the Tree of Life, and his artifacts, concrete and abstract, are indirectly products of the blind Darwinian processes in the same way spider webs and beaver dams are, so there is no *radical* discontinuity, no need for a skyhook, to get us from spiders and beaver dams to Turing and Turing machines. Still, there is a large gap to be filled, because Turing's way of making things was strikingly different from the spider's way and the

beaver's way, and we need a good evolutionary account of that difference. If competence *without* comprehension is so wonderfully fecund—capable of designing nightingales, after all—why do we need comprehension—capable of designing odes to nightingales and computers? Why and how did human-style comprehension arrive on the scene? First, let's make the contrast sharp and vivid.

If termites are impressive exemplars of competence without comprehension, capable of building strong, safe, air-conditioned homes without benefit of blueprints or bosses (the Queen termite is more like the Crown Jewels than a boss), Antoni Gaudí is a near-perfect model of the Intelligent Designer, a Godlike boss, armed from the outset with drawings and blueprints and manifestos full of passionately articulated reasons. His great church in Barcelona is an example of top-down creation that is hard to surpass, but Turing's original computer, the Pilot ACE (which can now be seen in the Science Museum in London), might beat it out for first prize. One of the first truly useful computers, it became operational in 1950 at the National Physical Laboratory in England, and it rivaled La Sagrada Familia in originality, intricacy—and cost. Both creators had to convince backers to fund their ambitious designs and both worked out elaborate diagrams, along with supporting explanations. So in each case, the eventual reality depended on the existence of prior representations, in the mind of a genius, of the purpose of the design, and hence the *raison d'être* of all the parts.[14] When it came to the actual construction of the artifacts, there were workers who were *relatively* uncomprehending, who had rather minimal appreciation of the point of their labors. Comprehension was distributed, of course: Gaudí didn't have to understand as much about how to mix mortar or carve stone as the masons on the job did, and Turing didn't have to be a virtuoso with a soldering gun or an expert on the techniques for manufacturing

14 Gaudí died in 1926 but left drawings and instructions and models that are still guiding the completion of the unfinished church; Turing left NPL before the Pilot ACE was completed, but he also left representations of the artifact to guide its completion.

vacuum tubes. Distribution of expertise or understanding of this sort is a hallmark of human creative projects, and it is clearly essential for today's high-tech artifacts, but not for all earlier artifacts. A lone artificer can make a spear, or even a kayak or a wooden wagon or a thatched hut, understanding every aspect of the design and construction, but not a radio or an automobile or a nuclear power plant.

A closer look at a few examples of human artifacts and the technology we have invented to make them will clarify the way-stations on the path from clueless bacteria to Bach, but first we need to introduce a term that began in philosophy and has been extended to several scientific and engineering enterprises.

Ontology and the manifest image

"Ontology" comes from the Greek word for *thing*. In philosophy, it refers to the set of "things" a person believes to exist, or the set of things defined by, or assumed by, some theory. What's in your ontology? Do you believe in ghosts? Then ghosts are in your ontology, along with tables and chairs and songs and vacations, and snow, and all the rest. It has proved more than convenient to extend the term "ontology" beyond this primary meaning and use it for the set of "things" that an animal can recognize and behave appropriately with regard to (whether or not animals can properly be said to have beliefs) and—more recently—the set of "things" a computer program has to be able to deal with to do its job (whether or not *it* can properly be said to have beliefs). Vacations are not in the ontology of a polar bear, but snow is, and so are seals. Snow is probably not in the ontology of a manatee, but outboard-motor propellers may well be, along with seaweed and fish and other manatees. The GPS system in your car handles one-way streets, left and right turns, speed limits, and the current velocity of your car (if it isn't zero, it may not let you put in a new target address), but its ontology also includes a number of satellites, as well as signals to and from those satellites, which it doesn't bother you with, but needs if it is to do its job.

The ontology of the GPS was intelligently designed by the programmers who built it, and the R&D process probably involved a lot of trial and error as different schemes were attempted and found wanting. The ontology of a polar bear or manatee was designed by some hard-to-sort-out combination of genetic evolution and individual experience. Manatees may have seaweed in their ontology the way human babies have nipples in theirs, instinctually, genetically designed over the eons. Any manatee with outboard-motor-propeller in its ontology has gained that from experience. We human beings have extremely varied ontologies. Some believe in witches and some believe in electrons and some believe in morphic resonances and abominable snowmen. But there is a huge common core of ontology that is shared by all normal human beings from quite an early age—six years old will capture almost all of it.

This common ontology was usefully named the *manifest image* by Wilfrid Sellars (1962). Consider the world we live in, full of other people, plants, and animals, furniture and houses and cars . . . and colors and rainbows and sunsets, and voices and haircuts, and home runs and dollars, and problems and opportunities and mistakes, among many other such things. These are the myriad "things" that are easy for us to recognize, point to, love or hate, and, in many cases, manipulate or even create. (We can't create sunsets, but in the right conditions we can create a rainbow with some water and a little ingenuity.) These are the things we use in our daily lives to anchor our interactions and conversations, and, to a rough approximation, for every noun in our everyday speech, there is a kind of thing it refers to. That's the sense in which the "image" is "manifest": it is obvious to all, and everybody knows that it is obvious to all, and everybody knows *that*, too. It comes along with your native language; it's the world according to *us*.[15] Sellars contrasted this

15 In fact, Sellars distinguished a "pre-scientific, uncritical, naïve conception of man-in-the-world . . . [which] might be called the 'original' image" (1962, p. 6ff) from what he called the manifest image, a "refinement or sophistication" of that original image. What he was mainly getting at in this distinction is that philosophers have been reflecting critically on the naïve conception for millennia, so the manifest image was not just folk metaphysics.

with the *scientific image*, which is populated with molecules, atoms, electrons, gravity, quarks, and who knows what else (dark energy, strings? branes?). Even scientists conduct most of their waking lives conceiving of what is going on in terms of the manifest image. ("Pass the pencil, please" is a typical bit of communication that depends on the manifest image, with its people and their needs and desires; their abilities to hear, see, understand, and act; the characteristic identifying marks of pencils, their size and weight, their use; and a host of other things. Making a robot that can understand and accede to such a request is far from trivial, unless you make a robot that can "understand" only that sentence and a few others.)

The scientific image is something you have to learn about in school, and most people (laypeople) acquire only a very cursory knowledge of it. These two versions of the world are quite distinct today, rather like two different species, but they were once merged or intertwined in a single ancestral world of "what everyone knows" that included all the local fauna and flora and weapons and tools and dwellings and social roles, but also goblins and gods and miasmas and spells that could jinx your life or guarantee your hunting success. Gradually our ancestors learned which "things" to oust from their ontologies and which new categories to introduce. Out went the witches, mermaids, and leprechauns, and in came the atoms, molecules, and germs. The early proto-scientific thinkers, such as Aristotle, Lucretius, and, much later, Galileo, conducted their inquiries without making a crisp distinction between the ontology of everyday life (the manifest image) and the ontology of science, but they were bold proposers of new types of things, and the most persuasive of these caught on. Undoing some of their most tempting mistakes, while creating the ontology of the scientific image, has been a major task of modern science.

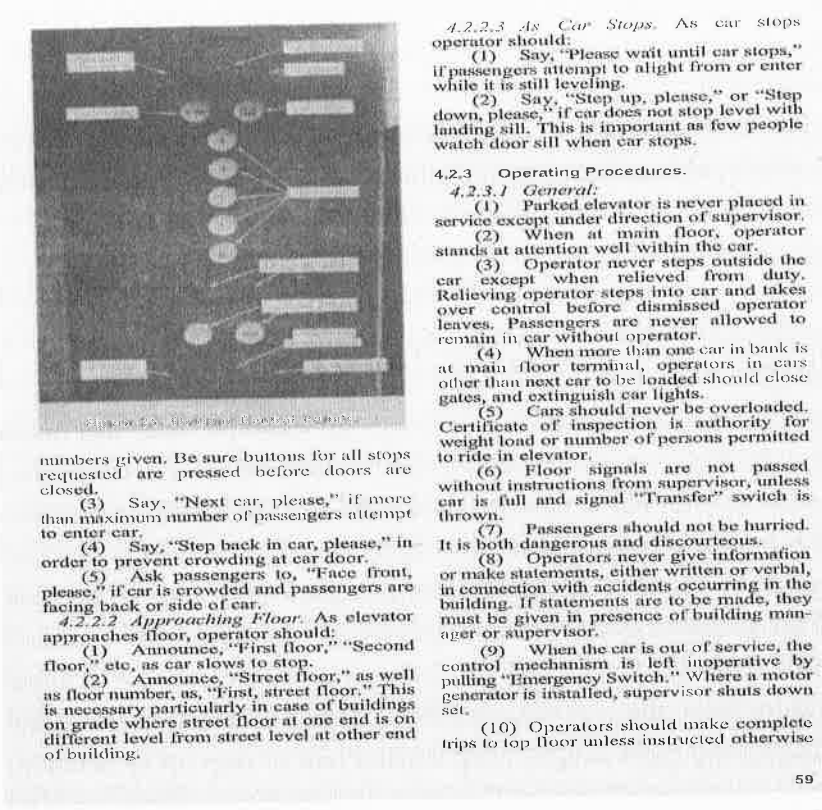
Unlike the term "ontology," "manifest image" and "scientific image" have not yet migrated from philosophy to other fields, but I'm doing my best to export them, since they have long seemed to me to be the best way I know to clarify the relationship between "our" world and the world of science. Where did the prescientific manifest

image come from? Sellars concentrated on the manifest image of human beings or societies. Should we extend the concept to other species? They have ontologies, in the extended sense. Do they also have manifest images, and how might they differ from ours? These questions are important to our inquiry because to understand what a great feat Darwin's strange inversion of reasoning was, we need to understand what Darwin was inverting and how *it* got that way.

Automating the elevator

It will help to start dead simple, with an example that has nothing to do with consciousness or even life: the electronic device that controls an automatic elevator. When I was a youngster, there were still human elevator operators, people whose job was to go up and down in an elevator all day, stopping at the right floors to take on and let off passengers. In the early days they manipulated a curious handle that could be swung clockwise or counterclockwise to make the elevator go up or down, and they needed skill to stop the elevator at just the right height. People often had to step up or down an inch or two on entering and leaving, and operators always warned people about this. They had lots of rules about what to say when, and which floors to go to first, and how to open the doors, and so forth. Their training consisted in memorizing the rules and then practicing: following the rules until it became second nature. The rules themselves had been hammered out over the years in a design process that made a host of slight revisions and improvements. Let's suppose that this process had more or less settled down, leaving an ideal rule book as its product. It worked wonderfully. Anybody who followed the rules exactly was an excellent elevator operator. (I located one of these antique rule books online, a US Army publication—not surprisingly, considering their pioneering role in drill and practice. Figure 4.1 reproduces a page.)

Now imagine what happened when a simple computer program could take over all the control tasks of the operator. (In fact, this happened gradually, with various automatic mechanical devices



numbers given. Be sure buttons for all stops requested are pressed before doors are closed.

(3) Say, "Next car, please," if more than maximum number of passengers attempt to enter car.

(4) Say, "Step back in car, please," in order to prevent crowding at car door.

(5) Ask passengers to, "Face front, please," if car is crowded and passengers are facing back or side of car.

4.2.2.2 Approaching Floor. As elevator approaches floor, operator should:

(1) Announce, "First floor," "Second floor," etc, as car slows to stop.

(2) Announce, "Street floor," as well as floor number, as, "First, street floor." This is necessary particularly in case of buildings on grade where street floor at one end is on different level from street level at other end of building.

4.2.2.3 As Car Stops. As car stops operator should:

(1) Say, "Please wait until car stops," if passengers attempt to alight from or enter while it is still leveling.

(2) Say, "Step up, please," or "Step down, please," if car does not stop level with landing sill. This is important as few people watch door sill when car stops.

4.2.3 Operating Procedures.

4.2.3.1 General:

(1) Parked elevator is never placed in service except under direction of supervisor.

(2) When at main floor, operator stands at attention well within the car.

(3) Operator never steps outside the car except when relieved from duty. Relieving operator steps into car and takes over control before dismissed operator leaves. Passengers are never allowed to remain in car without operator.

(4) When more than one car in bank is at main floor terminal, operators in cars other than next car to be loaded should close gates, and extinguish car lights.

(5) Cars should never be overloaded. Certificate of inspection is authority for weight load or number of persons permitted to ride in elevator.

(6) Floor signals are not passed without instructions from supervisor, unless car is full and signal "Transfer" switch is thrown.

(7) Passengers should not be hurried. It is both dangerous and discourteous.

(8) Operators never give information or make statements, either written or verbal, in connection with accidents occurring in the building. If statements are to be made, they must be given in presence of building manager or supervisor.

(9) When the car is out of service, the control mechanism is left inoperative by pulling "Emergency Switch." Where a motor generator is installed, supervisor shuts down set.

(10) Operators should make complete trips to top floor unless instructed otherwise

59

FIGURE 4.1: Elevator operator manual page.

being introduced to take the less skilled tasks away from the operator, but we'll imagine that elevators went from human operators to completely computer-controlled systems in one leap.) The elevator manufacturer, let's suppose, calls in a team of software engineers—programmers—and hands them the rule book that the human elevator operators have been following: "Here are *the specs*—this is a specification of the performance we want; make a computer program that follows all the rules in this book as well as the best human operators and we'll be satisfied." As the programmers go through the rule book, they make a list of all the actions that have to be taken, and the conditions under which they are prescribed or forbidden. In the process they can clean up some of the untidiness in the rule book. For instance, if they build in sensors to

ensure that the elevator always stops at exactly the right level, they can eliminate the loop that requires the operator to say, "Please step up" or "please step down," but they might leave in its place a simple (recorded) voice saying "[*n*]th floor; watch your step."

The rule book has instructions about how many passengers may be allowed on the elevator at any one time, and the programmers have to confront issues, such as do we install a turnstile so that the program can count people getting on and off? Probably not a good idea. A scale that weighs the occupants together is better, easier, and less intrusive. Look what that does to elevator ontology: instead of having a "count noun" like "passenger" or "occupant," it has a "mass noun" like "freight" or "cargo." We can say, metaphorically, that the elevator keeps asking itself "how much cargo?" not "how many passengers?" Similarly, we can note that the polar bear doesn't try to count snowflakes but is cognizant of the presence or absence of snow, or that the anteater slurps up a lot of ant on its tongue, unlike the insectivorous bird that tracks individual insects. And notice that just as we don't have to speculate about elevator *consciousness* to draw this distinction, we can treat the animals as having different ontologies without *settling* issues of whether they are conscious of their ontologies or simply the beneficiaries of designs that can be interpreted (by reverse engineers or forward engineers) as having those ontologies.

Back to elevator ontology. It may rely on "cargo" for some purposes, but it still needs to keep track of individual requests to which it must respond appropriately: "up," and "down," from outside; "five" or "ground floor" and "keep door open" from inside. And for safety it needs to self-monitor, to check its various organs periodically, to see if they are working correctly and actually in the state they are supposed to be in. It needs to light up buttons when they are pushed and turn the light off when the task ordered by the button is complete (or for other reasons). How conscientious (or obsessive-compulsive) the controller is can vary, but programs that are designed to be negligent about interference or failure will not make the grade for long. And if there are other elevators lined up in a common lobby (as in

a large office building or hotel), it will be important that the elevators communicate with each other, *or* that there is a master director that issues all the orders. (Designing the elevators to use "deictic" reference along the lines of "Where are you in relation to *where I am now*?" turns out to simplify and enhance the "cooperation" between individual elevators and eliminate the role of the omniscient master controller.)

It is useful to write the emerging control scheme in *pseudo-code*, a sort of mongrel language that is halfway between everyday human language and the more demanding system of source code. A line of pseudo-code might be along the lines of "if CALLFLOOR > CURRENTFLOOR, THEN ASCEND UNTIL CALLFLOOR = CURRENTFLOOR AND STOP; OPENDOOR. WAIT. . . ."

Once the plan is clear in pseudo-code and seems to be what is wanted, the pseudo-code can be translated into source code, which is a much more rigorous and structured system of operations, with definitions of terms—variables, subroutines, and so forth. Source code is still quite readily deciphered by human beings—after all, they write it—and hence the rules and terms of the rule book are still quite explicitly represented there, if you know how to look for them. This is made easier by two features: First, the names chosen for the variables and operations are usually chosen to wear their intended meaning on their sleeves (CALLFLOOR, WEIGHTSUM, TELLFLOOR . . .). Second, programmers can add *comments* to their source code, parenthetical explanations that tell other human readers of the source code what the programmer had in mind, and what the various parts are supposed to do. When you program, it is wise to add comments for yourself as you go, since you may easily forget what you thought the line of code was doing. When you go back to correct programming errors, these comments are very useful. Source code has to be carefully composed according to a strict syntax, with every element in the right place and all the punctuation in the right order since it has to be fed to a *compiler* program, which takes the source code and translates *it* into the sequences of fundamental operations that the actual machine (or virtual machine) can

execute. A compiler can't guess what a programmer means by a line of source code; the source code must tell the compiler exactly what operations to perform—but the compiler program may have lots of different ways of performing those tasks and will be able to figure out an efficient way under the circumstances.

Somewhere in the pseudo-code, amongst thousands of other statements, you will find a statement along the lines of

> IF WEIGHT-IN-POUNDS > *n* THEN STOP. OPEN DOOR.
>
> {Prevents elevator from moving if over maximum weight.
>
> After somebody steps out, reducing weight, normal operation resumes.}

The sentence in brackets is a *comment* that vanishes when the source code is compiled. Similarly, the capitalized terms don't survive in the code fed by the compiler to the computer chip that runs the program; they are also for the programmers, to help them remember which variable is which, and "IN-POUNDS" is in there to remind the programmers that the number they put in the program for maximum weight allowed better be in pounds. (In 1999, NASA's $125-million Mars Climate Orbiter got too close to Mars because one part of the control system was using meters and another part was using feet to represent the distance from the planet. The spacecraft got too close and destroyed itself. People make mistakes.) In short, the comments and labels help *us* understand the rationale of the design of the system but are ignored/invisible to the hardware. Once the program is finished and tested and deemed satisfactory, the compiled version can be burned into ROM (read-only-memory) where the CPU (central processing unit) can access it. The "rules" that were so explicit, so salient early in the design process, have become merely implicit in the zeroes and ones that get read by the hardware.

The point of this digression into elementary programming is that the finished working elevator has some interesting similarities to

living things yet also a profound difference. First, its activities are remarkably appropriate to its circumstances. It is a *good* elevator, making all the *right* moves. We might almost call it *clever* (like the best human elevator operators of yore). Second, this excellence is due to the fact that its design has the *right ontology*. It uses variables that keep track of all the features of the world that matter to getting its job done and is oblivious to everything else (whether the passengers are young or old, dead or alive, rich or poor, etc.). Third, it has *no need to know* what its ontology is, or why—the *rationale* of the program is something only the program's designers have to understand. They need to understand the rationale because of the nature of the R&D process that produced the finished program: it is a process of (quite) intelligent design. That is the profound difference we must clarify as we turn to the ontology of simple living things, products of evolution by natural selection, not intelligent design.

Even bacteria are good at staying alive, making the right moves, and keeping track of the things that matter most to them; and trees and mushrooms are equally clever, or, more precisely, cleverly designed to make the right moves at the right time. They all have elevator-type "minds," not elevated minds like ours.[16] They don't need minds like ours. And their elevator-minds are—must be—the products of an R&D process of trial and error that gradually structured their internal machinery to move from state to state in a way highly likely—not guaranteed—to serve their limited but vital interests. Unlike the elevator, their machinery was not designed by intelligent designers, who worked out, argued about, and thought about the rationales for the designs of the component pieces, so there is nothing—nothing at all, anywhere—that plays the roles of the labels or comments in a source code program. This is the key

16 Let me acknowledge that this claim is somewhat peremptory; I see no reason to believe that trees or bacteria have control systems that are more like our minds than elevator-control systems are, but I concede that it is *possible* that they do. I am treating this possibility as negligible, a non-zero strategic risk I am prepared to take.

to the transformation that Darwin and Turing achieved with their strange inversions of reasoning.

Elevators can do remarkably clever things, optimizing their trajectories, thereby saving time and energy, automatically adjusting their velocity to minimize discomfort of their passengers, "thinking of everything" that needs to be thought about, and obeying instructions and even answering the frequently asked questions. Good elevators earn their keep. They do this without any neurons, sense organs, dopamine, glutamate, or the other organic components of brains. So it seems fair to say that what they do so "cleverly" is a perfect case of competence without the slightest smidgen of comprehension or consciousness. Unless, of course, the machinery that provides them with this limited competence counts as having a smidgen, or maybe even two smidgens, of comprehension. (And in the same spirit, its prudent self-monitoring can be seen to be an elementary step towards consciousness.)

Whether or not we want to concede a minor, negligible touch of comprehension to the elevator, we should take the same line with bacteria, and with trees and mushrooms. They exhibit impressive competence at staying-alive-in-their-limited-niches, thanks to the well-designed machinery they carry with them, thanks to their genes. That machinery was designed by the R&D process of natural selection, however, so there is nothing anywhere at any time in that R&D history that represents the *rationales* of either the larger functions of whole systems or component functions of their parts the way comments and labels represent these functions for human designers. The rationales are nevertheless there to be discovered by reverse engineering. You can more or less count on it that there will be a *reason why* the parts are shaped as they are, why the behaviors are organized as they are, and that reason will "justify" the design (or have justified an earlier design that has now become either vestigial or transformed by further evolution to serve some newer function). The justification will be straightforward, in engineering terms: if you remove the element, or reshape the element, the system won't work, or won't work as well. Claims about such free-floating ratio-

nales should be, and can be, testable, and are confirmed beyond reasonable doubt in many cases.

Back to our elevator, successfully automated. *Tada!* One actual human being—not a figurative homunculus—has been replaced by a machine. And the machine *follows the same rules* as the human operator. Does it really? OK, it doesn't. It *sorta* follows the same rules. This is a nice intermediate case between a human being who memorizes—and hence literally represents in her mind, and consults—the rules that dictate her behavior, and the planets, whose orbits are elegantly described by equations that the planets "obey." We human beings also often occupy this intermediate level, when we have internalized or routinized through practice a set of explicit rules that we may then discard and even forget. (*i* before *e* except after *c*, or when it sounds like *a* as in "neighbor" and "weigh.") And it is also possible to sorta follow rules that have still not been made explicit: the rules of English grammar, for instance, which continue to challenge linguists. Put in terms of this example, linguists today are still thrashing around trying to write a satisfactory version of the "rule book" for speaking English, while every ten-year-old native English speaker has somehow installed and debugged a pretty good version of the executable object code for the control task of speaking and understanding the language.

Before we take up the minds of animals, I want to turn to some further examples of the design of artifacts that will help isolate the problem that evolution solved when it designed competent animals.

The intelligent designers of Oak Ridge and GOFAI

After seventy years there are still secrets about World War II that have yet to emerge. The heroic achievements of Alan Turing in breaking the German Enigma code at Bletchley Park are now properly celebrated even while some of the details are still considered too sensitive to make public. Only students of the history of atomic energy engineering are likely to be well acquainted with the role

that General Leslie Groves played in bringing the Manhattan Project to a successful conclusion. It took only six years from the day in August of 1939 when the Einstein-Szilard letter arrived on President Roosevelt's desk informing him of the prospect of an atomic bomb until the dropping of the first bomb on Hiroshima on August 6, 1945. The first three years went into basic research and "proof of concept," and almost everybody involved in those early years knew exactly what they were trying to accomplish. In 1942, Leslie Groves was appointed director of what came to be called the Manhattan Project, and in three incredibly pressured years intertwining further R&D with the colossal (and brand new) task of refining weapons grade uranium, thousands of workers were recruited, trained, and put to work, mostly controlling the newly invented machines for separating out the isotope uranium 235, which was a fraction of 1% of the previously refined uranium 238.

At the height of operations over 130,000 people worked full time on the project, and only a tiny percentage of them had any idea at all what they were making. Talk about competence without comprehension! The Need to Know principle was enforced to the maximum degree. In K-25, the gaseous diffusion plant in the instant city of Oak Ridge, Tennessee, tens of thousands of men and women worked around the clock attending to dials and pushing buttons and levers on a task that they came to perform with true expertise and no understanding at all. As their reactions in the aftermath of Hiroshima made clear, they didn't know if they were making airplane parts or crankcase oil for submarines or what. Think of the planning required to create a training system that could turn them into experts without letting on what kind of experts they were. The level of secrecy was higher than ever before (or since, probably). Leslie Groves and the planners all needed to know a great deal about the project, of course; they were intelligent designers, armed with detailed and precise understanding of the specs of the task; only by using that understanding could they create the sheltered environment for uncomprehending competence.

The project proved one point beyond a shadow of a doubt: it is

possible to create very reliable levels of high competence with almost no comprehension for rather insulated tasks. So far as I can determine, to this day the precise distribution of understanding through the entire workforce of the Manhattan Project is a closely guarded secret. What did the engineers and architects who designed the K-25 building need to know? It was the largest building in the world when they constructed it in a matter of months. Some of them obviously needed to know what the miles of highly specialized pipes were going to be used for, but probably the designers of the roof, the foundation, and the doors had no inkling. It is pleasant to reflect that while Groves and his team were intelligently designing a system of thousands of human jobs that required minimal comprehension, Turing and his team on the other side of the Atlantic were intelligently designing a system that could replace those clueless homunculi with electronics. A few years later, scientists and engineers, most of whom had contributed to one or another of these pathbreaking wartime projects, began exploiting Turing's invention of uncomprehending competent building blocks to create the audacious field of Artificial Intelligence.

Turing himself (1950) prophesied that by the turn of the century "the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted." The early work in the field was brilliant, opportunistic, naïvely optimistic, and, one might well say, saturated with hubris. An artificial intelligence ought to be able to see, as well as think, certainly, so let's first design a seeing machine. The notorious "summer vision project" of the mid-1960s at MIT was an attempt to "solve vision" over one long vacation, leaving harder problems for later! By today's standards the "giant electronic brains" on which the early work was conducted were tiny and achingly slow, and one of the side effects of these limitations was that efficiency was a high-priority goal. Nobody would bother creating a computer model that would take days to respond to a realistic set of inputs, especially if the goal was to harness the computer's unprecedented speed to handle real-world, real-time problems.

Early AI, or GOFAI (Good Old-Fashioned AI [Haugeland 1985]), was a "top-down," "intellectualist" approach to Artificial Intelligence: write down what human experts know, in a language the computer can manipulate with *inference engines* that could patrol the "huge" memory banks stocked with this carefully handcrafted *world knowledge*, deducing the theorems that would be needed to make informed decisions and control appropriately whatever limbs or other effectors the intelligence had. GOFAI can be seen in retrospect to have been an exercise in creating something rather Cartesian, a rationalistic expert with myriads of *propositions* stored in its memory, and all the *understanding* incorporated in its ability to draw conclusions from the relevant axioms and detect contradictions in its world knowledge—as efficiently as possible. What is an intelligent agent after all, but a well-informed rational being, which can think fast enough, using the propositions it knows, to plan actions to meet whatever contingencies arise? It seemed like a good idea at the time, and to some researchers in the field, it still does.[17]

The premium on speed and efficiency dictated working first on "toy problems." Many of these ingeniously scaled-down problems were more or less solved, and the solutions have found applications in the not particularly demanding world of controllers in restricted environments (from elevators and dishwashers to oil refineries and airplanes), and in medical diagnosis, game playing, and other carefully circumscribed areas of investigation or interaction: making airline reservations, spell-checking and even grammar checking, and the like. We can think of these designs as rather indirect descendants of the heavily insulated systems created by Groves and his elite team of intelligent designers, adhering to the Need to Know principle, and relying on the comprehension of the *designers* to contrive systems composed of subsystems that were foresightedly equipped with exactly the competences they would need in order to handle

17 Douglas Lenat's CYC project is the ongoing attempt to create such an artificial intelligence, and after thirty years of work by hundreds of coders (known as CYClists), it has over a million hand-defined concepts in its memory.

the problems they might face. Since for all their brilliance the early AI designers weren't omniscient (and time was of the essence), they restricted the range and variety of inputs each subsystem had to accept, and deal with, creating programs harboring thousands of sheltered workshops to protect the idiot savants (subroutines) that worked there.

Much was learned, and a lot of good practices and techniques were invented and refined, but they mainly helped dramatize how truly difficult the task of designing a free-wheeling, imaginative, open-ended human mind is. The dream of a hand-coded, top-down-organized, bureaucratically efficient know-it-all, a walking (or at least talking) encyclopedia, is not yet entirely extinguished, but as the size of the project became clearer, there has been a salutary shift of attention to a different strategy: using colossal amounts of Big Data and the new statistical pattern-finding techniques of data-mining and "deep learning" to eke out the necessary information in a more bottom-up way.

I will have much more to say about these developments later; for the time being, the point that we need to recognize is that the vast increase in speed and size of computers over the years has opened up the prospect of exploring "wasteful," "mindless," less "bureaucratic," more evolution-like processes of information extraction, and these are achieving impressive results. Thanks to these new perspectives, we can now think in some detail about the question of how the *relatively* simple systems that control bacteria, worms, and termites, for example, might have evolved by the bottom-up, foresightless, brute force processes of natural selection. In other words, we want to see how evolution might play the Leslie Groves role in organizing clueless operatives into effective teams *without* the luxury of Groves's understanding and foresight.

Top-down intelligent designing works. The policy of planning ahead, articulating the problems, refining the tasks, and clearly representing the reasons for each step is a strategy that has not just seemed obvious to inventors and problem-solvers for millennia; it has proven itself in countless triumphs of foresight and ingenuity

in every field of human endeavor: from science and engineering to political campaigns and cooking, farming, and navigation. Before Darwin, it was seen as the only way design could be accomplished; design without an intelligent designer was deemed impossible. But top-down design is in fact responsible for much less of the design in our world than is commonly appreciated, and for some of the "achievements of creative skill," to echo Beverley once again, victory has so far eluded it. Darwin's "strange inversion of reasoning" and Turing's equally revolutionary inversion were aspects of a single discovery: competence without comprehension. Comprehension, far from being a Godlike talent from which all design must flow, is an emergent effect of systems of uncomprehending competence: natural selection on the one hand, and mindless computation on the other. These twin ideas have been proven beyond a reasonable doubt, but they still provoke dismay and disbelief in some quarters, which I have tried to dispel in this chapter. Creationists are not going to find commented code in the inner workings of organisms, and Cartesians are not going to find an immaterial *res cogitans* "where all the understanding happens."

5

The Evolution of Understanding

Animals designed to deal with affordances

Animals are designed by natural selection, of course, but such a declaration of confidence in evolution is not informative. How, more particularly, might evolution turn this trick? One of the fruits of our interlude on the designing of an elevator controller and its artifactual kin is a sharper sense of how different that R&D process is from evolution by natural selection. The computer on which the designers—the programmers—test and run their solutions is itself a product of intelligent design, as we have noted, and its initial set of building-block competences—arithmetic and conditional branching—invite all would-be programmers to conceive of their tasks in the top-down way as well, as *problem-solving* in which they try to embody *their* understanding of the problem in the solutions they build.

"How else?" one might well ask. Intelligent design of this sort starts with a goal (which may well be refined or even abandoned along the way) and works top-down, with the designers using everything they know to guide their search for solutions to the design problems (and sub-problems, and sub-sub-problems . . .) they set for themselves. Evolution, in contrast, has no goals, no predefined

problems, and no comprehension to bring to the task; it myopically and undirectedly muddles along with what it has already created, mindlessly trying out tweaks and variations, and keeping those that prove useful, or at least not significantly harmful.

Could something as intellectually sophisticated as a digital computer, for instance, ever evolve by bottom-up natural selection? This is very hard to imagine or even to take seriously, and this has inspired some thinkers to conclude that since evolution couldn't create a computer (or a computer program to run on it), human minds must not be products of natural selection alone, and the aspirations of Artificial Intelligence must be forlorn. The mathematician and physicist Roger Penrose (1989) is the most illustrious example. For the sake of argument let's concede that evolution by natural selection could not *directly* evolve a living digital computer (a Turing machine *tree* or a Turing machine *turtle*, for example). But there is an indirect way: let natural selection first evolve human minds, and then *they* can intelligently design *Hamlet*, La Sagrada Familia, and the computer, among many other wonders. This bootstrapping process seems almost magical at first, even self-contradictory. Isn't Shakespeare, or Gaudí, or Turing a more magnificent, brilliant "creation" than any of their brainchildren? In some regards, yes, of course, but it is also true that their brainchildren have features that couldn't come into existence without them.

If you landed on a distant planet and were hunting along its seashore for signs of life, which would excite you more, a clam or a clam rake? The clam has billions of intricate moving parts, while the clam rake has just two crude, fixed parts, but it must be an artifact of some living thing, something much, much more impressive than a clam. *How could a slow, mindless process build a thing that could build a thing that a slow mindless process couldn't build on its own?* If this question seems to you to be unanswerable, a rhetorical question only, you are still in thrall to the spell Darwin broke, still unable to adopt Darwin's "strange inversion of reasoning." Now we can see how strange and radical it is: a process with

no Intelligent Designer can create intelligent designers who can then design things that permit us to understand how a process with no Intelligent Designer can create intelligent designers who can then design things.

The intermediate steps are instructive. What about the clam rake gives away its artifactual status? Its very simplicity, which indicates its dependence on something else for its ability to defy the Second Law of Thermodynamics, persisting as uniform and symmetrical collections of atoms of elements in improbable juxtapositions. Something gathered and refined these collections. Something complicated.

Let's return once more to simple organisms. The idea that every organism has its ontology (in the elevator sense) was prefigured in Jakob von Uexküll's (1934) concept of the organism's *Umwelt*,

FIGURE 5.1: Clam rake. © *Daniel C. Dennett.*

the behavioral environment that consists of all the things that matter to its well-being. A close kin to this idea is the psychologist J. J. Gibson's (1979) concept of *affordances*: "What the environment offers the animal for good or ill." Affordances are the relevant opportunities in the environment of any organism: things to eat or mate with, openings to walk through or look out of, holes to hide in, things to stand on, and so forth. Both von Uexküll and Gibson were silent about the issue of whether consciousness (in some still-to-be-defined sense) was involved in having an *Umwelt* populated by affordances, but since von Uexküll's case studies included amoebas, jellyfish, and ticks, it is clear that he, like Gibson, was more interested in characterizing the *problems faced and solved* by organisms than on how, internally, these solutions were carried out. The sun is in the ontology of a honey bee; its nervous system is designed to exploit the position of the sun in its activities. Amoebas and sunflowers also include the sun in their *Umwelten*; lacking nervous systems, they use alternative machinery to respond appropriately to its position. So the engineer's concept of elevator ontology is just what we need at the outset. We can leave until later the questions of whether, when, and why the ontology of an organism, or a lineage of organisms, becomes *manifest* in consciousness of some sort and not just *implicit* in the designed responses of its inner machinery. In other words, organisms can be the beneficiaries of design features that imply ontologies without themselves *representing* those ontologies (consciously, semiconsciously, or unconsciously) in any stronger sense. The shape of a bird's beak, together with a few other ancillary features of anatomy, *imply* a diet of hard seeds, or insects or fish, so we can stock the *Umwelten* of different species of birds with hard seeds, insects, and fish, as species-specific *affordances* on the basis of these anatomical features alone, though of course it is wise to corroborate the implication by studying behavior if it is available. The shape of the beak does not in any interesting sense *represent* its favored foodstuff or way of obtaining it.

Paleontologists draw conclusions about the predatory prefer-

ences and other behaviors of extinct species using this form of inference, and it is seldom noted that it depends, ineliminably, on making adaptationist assumptions about the designs of the fossilized creatures. Consider Niles Eldredge's (1983) example of Fisher's (1975) research on horseshoe crab swimming speed. He cites it to demonstrate that asking the historical question "what has happened?" ("how come") is a better tactic than asking the adaptationist question ("what for"), with its optimality assumptions. But Fisher's conclusion about how fast the ancient horseshoe crabs swam

> depends on a very safe adaptationist assumption about what is good: *Faster is better—within limits.* The conclusion that Jurassic horseshoe crabs swam faster depends on the premise that they would achieve maximal speed, given their shape, by swimming at a certain angle, *and* that they would swim so as to achieve maximal speed. So . . . [Fisher needs an] entirely uncontroversial, indeed tacit, use of optimality considerations to get *any purchase at all* on "what happened" 150 million years ago. (Dennett 1983)

Remember, biology is reverse engineering, and reverse engineering is methodologically committed to optimality considerations. "What is—or was—this feature *good for*?" is always on the tip of the tongue; without it, reverse engineering dissolves into bafflement.

As I said in the opening paragraph of the book, bacteria don't know they are bacteria, but of course they respond to other bacteria in bacteria-appropriate ways and are capable of avoiding or tracking or trailing things they distinguish in their *Umwelt*, without needing to have any idea about what they are doing. Bacteria are in the ontology of bacteria the same way floors and doors are in the ontology of elevators, only bacteria are much more complicated. Just as there are reasons why the elevator's control circuits are designed the way they are, there are reasons why the bacteria's internal protein control networks are designed the way they are: in both cases the designs have been optimized to handle the problems encountered efficiently and effectively.[18] The chief difference is that the design of the elevator circuits was done by intelligent designers who had worked out descriptions of the problems, and representations of *reasoned* solutions, complete with justifications. In the R&D history of the bacteria, there was no source code, and no comments were ever composed, to provide hints of what Mother Nature intended. This does not stop evolutionary biologists from assigning functions to some evolved features (webbed feet are for propulsion in water), and interpreting other features as mistakes of Nature (a two-headed calf). Similarly, literary editors of texts of long-dead authors don't have to rely on autobiographical divulgences left behind in the author's papers to interpret some unlikely passages as deliberately misleading and others as typographical errors or memory lapses.

Software development is a relatively new domain of human endeavor. While still in its infancy many foibles and glitches have been identified and corrected, and a Babel Tower of new programming languages has been created, along with a host of software-writing tools to make the job easier. Still, programming is an "art," and even commercially released software from the best purveyors always turns out to have "bugs" in it that require correction in post-purchase updates. Why hasn't debugging been automated, eliminating these costly errors from the outset? The most intelligent human designers, deeply informed about the purposes of the software, still find debugging code a daunting task, even when they can examine carefully commented source code produced under strictly regimented best practices (Smith 1985, 2014). There is a reason why debugging cannot be completely automated: what counts as a bug depends on all the purposes (and sub-purposes, and sub-sub-purposes) of the

18 There is much controversy about using the term "optimize" when referring to the "good enough" products of natural selection. The process of natural selection cannot "consider all things" and is always in the midst of redesign, so it is not guaranteed to find the *optimal* solution to any specific design problem posed, but it does amazingly well, typically better than intelligent human designers who are striving for optimal design.

software, and specifying in sufficient detail what those purposes are (in order to feed them to one's imagined automated debugger program) is, at least for practical purposes, the very same task as writing debugged code in the first place![19] Writing and debugging computer code for ambitious systems is one of the most severe tests of human imagination yet devised, and no sooner does a brilliant programmer devise a new tool that relieves the coder of some of the drudgery than the bar is raised for what we expect the coder to create (and test). This is not an unprecedented phenomenon in human activity; music, poetry, and the other arts have always confronted the would-be creator with open-ended spaces of possible "moves" that did not diminish once musical notation, writing, and ready-made paints were made available, nor does artistic creation become routinized by the addition of synthesizers and MIDI files, word-processing and spell-checking, and million-color, high-resolution computer graphics.

How does Nature debug its designs? Since there is no source code or comments to read, there can be no debugging by brilliant intellectual explanation; design revision in Nature must follow the profligate method of releasing and test-driving many variants and letting the losers die, *unexamined*. This won't necessarily find the globally optimal design but the best locally accessible versions will thrive, and further test-driving will winnow the winners further, raising the bar slightly for the next generation.[20] Evolution is, as Richard Dawkins's (1986) memorable title emphasizes, the Blind Watchmaker, and given the R&D method used, it is no wonder that evolution's products are full of opportunistic, short-sighted, but deviously effective twists and turns—effective except when they

19 Legendary software designer Charles Simonyi, the principal creator of Microsoft Word, has devoted more than twenty years to the task of creating what he calls "Intentional Software," which would ideally solve this problem or a valuable subset of these problems. The fact that several decades of high-quality work by a team of software engineers has not yet yielded a product says a lot about the difficulty of the problem.

20 Evolution explores the "adjacent possible," see Kauffman (2003).

aren't! One of the hallmarks of design by natural selection is that it is full of bugs, in the computer programmer's sense: design flaws that show up only under highly improbable conditions, conditions never encountered in the finite course of R&D that led to the design to date, and hence not yet patched or worked around by generations of tinkering. Biologists are very good at subjecting the systems they are studying to highly improbable conditions, imposing extreme challenges to see where and when the systems fail, *and why*.

What they typically discover, when reverse engineering an organism, is like the all-but-undecipherable "spaghetti code" of undisciplined programmers. If we make the effort to decipher spaghetti code, we can usually note which unlikely possibilities *never occurred to* the designers in their myopic search for the best solution to the problems posed for them. *What were they thinking?* When we ask the same question about Mother Nature, the answer is always the same: nothing. No thinking was involved, but nevertheless she muddled through, cobbling together a design so effective that it has survived to this day, beating out the competition in a demanding world until some clever biologist comes along and exposes the foibles.

Consider *supernormal stimuli*, a design glitch found in many organisms. Niko Tinbergen's (1948, 1951, 1953, 1959) experiments with seagulls revealed a curious bias in their perceptual/behavioral machinery. The adult female has an orange spot on her beak, at which her chicks instinctually peck, to stimulate their mother to regurgitate and feed them. What if the orange spot were bigger or smaller, brighter or less distinct? Tinbergen showed that chicks would peck even more readily at exaggerated cardboard models of the orange spot, that supernormal stimuli evoked supernormal behaviors. Tinbergen also showed that birds that laid light blue, gray-dappled eggs preferred to sit on a bright blue black polka-dotted fake egg so large that they slid off it repeatedly.

"This isn't a bug, it's a feature!" is the famous programmers' retort, and the case can be made for supernormal stimuli. As long as their *Umwelt* doesn't have sneaky biologists with vivid imaginations challenging the birds with artificial devices, the system works

very well, focusing the organism's behavior on what (almost always) matters. The free-floating rationale of the whole system is clearly good enough for practical purposes, so Mother Nature was wise not to splurge on something more foolproof that would detect the ruse. This "design philosophy" is everywhere in Nature, providing the opportunities for arms races in which one species exploits a shortcut in another species' design, provoking a counter-ploy in Design Space that ratchets both species to develop ever better defenses and offenses. Female fireflies sit on the ground watching male fireflies emit patterns of flashes, showing off and hoping for an answer from the female. When the female makes her choice and flashes back, the male rushes down for a mating. But this ingenious speed-dating system has been invaded by another species of firefly, Photuris, that pretends to be a female, luring the males to their death. The Photuris prefers males with longer, stronger signals, so the males are evolving shorter love letters (Lewis and Cratsley 2008).

Higher animals as intentional systems: the emergence of comprehension

Competence without comprehension is Nature's way, both in its methods of R&D and in its smallest, simplest products, the brilliantly designed motor proteins, proofreading enzymes, antibodies, and the cells they animate. What about multicellular organisms? When does comprehension emerge? Plants, from tiny weeds to giant redwood trees, exhibit many apparently clever competences, tricking insects, birds, and other animals into helping them reproduce, forming useful alliances with symbionts, detecting precious water sources, tracking the sun, and protecting themselves from various predators (herbivores and parasites). It has even been argued (see, e.g., Kobayashi and Yamamura 2003; Halitschke et al. 2008) that some species of plants can warn nearby kin of impending predation by wafting distress signals downwind when attacked, permitting those that receive the signals to heighten their defense mechanisms in anticipation, raising their

toxicity or generating odors that either repel the predators or lure symbionts that repel the predators. These responses unfold so slowly that they are hard to see as proper behaviors without the benefit of time-lapse photography, but, like the microscopic behaviors of single cells, they have clear rationales that need not be understood by the actors.

Here we see emerging something like a double standard of attribution. It is well-nigh impossible to describe and explain these organized-processes-in-time without calling them behaviors and *explaining* them the way we explain our own behaviors, by citing reasons and assuming that they are guided by something like perceptual monitoring, the intake of information that triggers, modulates, and terminates the responses. And when we do this, we *seem* to be attributing not just competence but also the comprehension that—in us—"normally goes with" such behavioral competence. We are anthropomorphizing the plants and the bacteria in order to understand them. This is not an intellectual sin. We are *right* to call their actions behaviors, to attribute these competences to the organisms, to explain their existence by citing the rationales that account for the benefits derived from these competences by the organisms in their "struggle" for survival. We are right, I am saying, to adopt what I call the intentional stance. The only mistake lies in attributing *comprehension* to the organism or to its parts. In the case of plants and microbes, fortunately, common sense intervenes to block that attribution. It is easy enough to understand how their competence can be provided by the machinery without any *mentality* intruding at all.

Let's say that organisms that have spectacular competences without any need for comprehension of their rationales are *gifted*. They are the beneficiaries of talents bestowed on them, and these talents are not products of their own individual investigation and practice. You might even say they are *blessed* with these *gifts*, not from God, of course, but from evolution by natural selection. If our imaginations need a crutch, we can rely on the obsolescing stereotype of the robot as a mindless mechanism: plants don't have understanding; they're living *robots*. (Here's a prediction: in a hundred years, this

will be seen as an amusing fossil of *biocentrism*, a bit of prejudice against comprehending robots that survived well into the twenty-first century.)

While we're on this topic, it's interesting to recall that in the twentieth century one of the most popular objections to GOFAI was this:

> The so-called intelligence in these programs is really just the intelligence—the understanding—of the programmers. The programs don't understand anything!

I am adopting and adapting that theme but not granting understanding (yet) to anyone or anything:

> The so-called intelligence in trees and sponges and insects is not theirs; they are just brilliantly designed to make smart moves at the right time, and while the design is brilliant, the designer is as uncomprehending as they are.

The opponents of GOFAI thought they were stating the obvious when they issued their critique of so-called intelligent machines, but see how the emotional tug reverses when the same observation is ventured about animals. Whereas—I surmise—most readers will be quite comfortable with my observation that plants and microbes are merely gifted, blessed with well-designed competences, but otherwise clueless, when I then venture the same opinion about "higher" animals, I'm being an awful meanie, a killjoy.

When we turn to animals—especially "higher" animals such as mammals and birds—the temptation to attribute comprehension in the course of describing and explaining the competences is much greater, and—many will insist—entirely appropriate. Animals really do understand what they're doing. See how amazingly clever they are! Well, now that we have the concept of competence without comprehension firmly in hand, we need to reconsider this gracious opinion. The total weight of all life on the planet now—the biomass—is currently estimated as more than half made up of bacteria and

other unicellular "robots," with "robotic" plants making up more than half the rest. Then there are the insects, including all the clueless termites and ants that outweigh the huge human population celebrated by MacCready. We and our domesticated animals may compose 98% of the *terrestrial vertebrate* biomass, but that is a small portion of life on the planet. Competence *without* comprehension is the way of life of the vast majority of living things on the planet and should be the default presumption until we can demonstrate that some individual organisms really do, in one sense or another, *understand* what they are doing. Then the question becomes when, and why, does the design of organisms start *representing* (or otherwise intelligently incorporating) the free-floating rationales of their survival machinery? We need to reform our imaginations on this issue, since the common practice is to *assume* that there is some kind of understanding in "higher animals" wherever there is a rationale.

Consider a particularly striking example. Elizabeth Marshall Thomas is a knowledgeable and insightful observer of animals (including human animals), and in one of her books, *The Hidden Life of Dogs* (1993), she permits herself to imagine that dogs enjoy a wise understanding of their ways: "For reasons known to dogs but not to us, many dog mothers won't mate with their sons" (p. 76). There is no doubt about their instinctive resistance to such inbreeding; they probably rely mainly on scent as their cue, but who knows what else contributes—a topic for future research. But the suggestion that dogs have any more insight into the reasons for their instinctual behaviors and dispositions than we have into ours is romanticism run wild. I'm sure she knows better; my point is that this lapse came naturally to her, an extension of the prevailing assumption, not a bold proposal about the particular self-knowledge of dogs. This is like a Martian anthropologist writing, "For reasons known to human beings but not to us, many human beings yawn when sleepy and raise their eyebrows when they see an acquaintance." There *are* reasons for these behaviors, *what for* reasons, but they are not *our* reasons. You *may* fake a yawn or raise your eyebrows for a reason—to give a deliberate signal, or to feign acquaintance with an

attractive but unfamiliar person you encounter—but in the normal case you don't even realize you do it, and hence have no occasion to know why you do it. (We still don't know why we yawn—and certainly dogs aren't ahead of us on this point of inquiry, though they yawn just as we do.)

What about more obviously deliberate behaviors in animals? Cuckoos are *brood parasites* that don't make their own nests. Instead, the female cuckoo surreptitiously lays her egg in the nest of a host pair of some other species of birds, where it awaits the attentions of its unwittingly adoptive parents. Often, the female cuckoo will roll one of the host eggs out of the nest—in case the host parents can count. And as soon as the cuckoo chick is hatched (and it tends to incubate more quickly than the host eggs), the little bird goes to great efforts to roll any remaining eggs out of the nest. Why? To maximize the nurture it will get from its adoptive parents. The video clips of this behavior by the hatchling cuckoo are chilling demonstrations of efficient, competent killing, but there is no reason to suppose that mens rea (guilty intention, in the law) is in place. The baby bird knows not what it is doing but is nevertheless the beneficiary of its behavior. What about nest building in less larcenous species? Watching a bird build a nest is a fascinating experience, and there is no doubt that highly skilled weaving and even sewing actions are involved (Hansell 2000). There is quality control, and a modicum of learning. Birds hatched in captivity, never having seen a nest being built, will build a serviceable species-typical nest out of the available materials when it is time to build a nest, so the behavior is instinctual, but it will build a better one the next season.

How much understanding does the nest-building bird have? This can be, and is being, probed by researchers (Hansell 2000, 2005, 2007; Walsh et al. 2011; Bailey et al. 2015), who vary the available materials and otherwise interfere with the conditions to see how versatile and even foresighted the birds can be. Bearing in mind that evolution can only provide for challenges encountered during R&D, we can predict that the more novel the artificial intrusions in the bird's *Umwelt* are, the less likely it is that the bird will interpret

them appropriately, *unless* the bird's lineage evolved in a highly varied selective environment that obliged natural selection to settle on designs that are not entirely *hard-wired* but have a high degree of plasticity and the learning mechanisms to go with it. Interestingly, when there isn't enough stability over time in the selective environment to permit natural selection to "predict" the future accurately (when "selecting" the best designs for the next generation), natural selection does better by leaving the next generation's design partially unfixed, like a laptop that can be configured in many different ways, depending on the purchaser's preferences and habits.[21] Learning can take over where natural selection left off, optimizing the individuals in their own lifetimes by extracting information from the world encountered and using it to make local improvements. We will soon turn to a closer examination of this path to understanding, but, first, I want to explore a few more examples of behaviors with free-floating rationales and their implications.

You may have seen video of antelopes being chased across the plains by a predator and noticed that some of the antelopes leap high in the air during their attempts to escape their pursuer. This is called stotting. Why do antelopes stot? It is clearly beneficial, because antelopes that stot seldom get caught and eaten. This is a causal regularity that has been carefully observed, and it demands a *what for* explanation. No account of the actions of all the proteins and the like in the cells of all the antelopes and predators chasing them could reveal why this regularity exists. For an answer we need the branch of evolutionary theory known as costly signaling theory (Zahavi 1975; Fitzgibbon and Fanshawe 1988). The strongest and fastest of the antelopes stot in order to advertise their fitness to the pursuer, signaling, in effect, "Don't bother chasing me; I'm too hard to catch; concentrate on one of my cousins who isn't able

21 How can I speak of evolution, which famously has no foresight, being able or unable to *predict* anything? We can cash out this handy use of the intentional stance applied to evolution itself by saying, less memorably and instructively, that highly variable environments *have no information* about future environments for natural selection to (mindlessly) exploit (see chapter 6).

to stot—a much easier meal!" and the pursuer takes this to be an honest, hard-to-fake signal and ignores the stotter. This is both an act of *communication* and an act with only a free-floating rationale, which need not be appreciated by either antelope or lion. That is, the antelope may be entirely oblivious of why it is a good idea to stot if you can, and the lion may not understand why it finds stotting antelopes relatively unattractive prey, but if the signaling wasn't honest, costly signaling, it couldn't persist in the evolutionary arms race between predator and prey. (If evolution tried a "cheap" signal, like tail flicking, which every antelope, no matter how frail or lame, could send, it wouldn't pay lions to attend to it, so they wouldn't.) This may seem an overly skeptical *killjoy* demotion of the intelligence of both antelope and lion, but it is the strict application of the same principles of reverse engineering that can account for the cuckoo and the termite and the bacterium. The rule of attribution must be then, if the competence observed can be explained without appeal to comprehension, don't indulge in extravagant anthropomorphism. Attributing comprehension must be supported by demonstrations of much more intelligent behavior. Since stotting is not (apparently) an element in a more elaborate system of interspecies or intraspecies communication on many topics, the chances of finding a need for anything that looks like comprehension here are minimal. If you find this verdict too skeptical, try to imagine some experiments that could prove you right.

How could experiments support the verdict of comprehension? By showing that the animals can do what we comprehenders can do with variations on the behavior. Stotting is a variety of showing off, or bragging, and we can do that, but we also can bluff or refrain from bragging or showing off if conditions arise that render such behavior counterproductive or worse. We can modulate our bragging, tuning it to different audiences, or do some transparently exaggerated bragging to telegraph that we don't really mean it and are making a joke. And so forth, indefinitely. Can the antelope do any of this? Can it refrain from stotting in circumstances that, in novel ways, make stotting inappropriate? If so, this is some evidence

that it has—and uses—some minimal understanding of the rationale of its actions.

A rather different free-floating rationale governs the injury-feigning, ground-nesting bird, such as a piping plover, that lures a predator away from her nest by seeming to have a broken wing, keeping just out of the predator's reach until she has drawn it far from her nest. Such a "distraction display" is found in many very widely separated species of ground-nesting birds (Simmons 1952; Skutch 1976). This seems to be *deception* on the bird's part, and it is commonly called that. Its purpose is to *fool* the predator. Adopting Dawkins's (1976) useful expository tactic of inventing "soliloquies," we can devise a soliloquy for the piping plover:

> I'm a low-nesting bird, whose chicks are not protectable against a predator that discovers them. This approaching predator can be *expected* soon to discover them unless I distract it; it could be distracted by its *desire* to catch and eat me, but only if it *thought* there was a *reasonable* chance of its actually catching me (it's no dummy); it would contract just that *belief if I gave it evidence that* I couldn't fly anymore; I could do that by feigning a broken wing, and so on.

Talk about sophistication! Not just a goal, but also a *belief* about an *expectation*, and a *hypothesis* about the *rationality* of the predator and a *plan* based on that hypothesis. It is unlikely in the extreme that any feathered "deceiver" is capable of such mental representation. A more realistic soliloquy to represent what is "in the mind" of the bird would be something like: "Here comes a predator; all of a sudden I feel this tremendous urge to do that silly broken-wing dance. I wonder why?" But even this imputes more *reflective* capacity to the bird than we have any warrant for. Like the elevator, the bird has been designed to make some important discriminations and do the right thing at the right time. Early investigators, rightly deeming the sophisticated soliloquy too good to be true as an account of the bird's *thinking*, were tempted to hypothesize that the behavior

was not deliberate at all, but a sort of panic attack, composed of unguided spasms that had the beneficial side effect of attracting the attention of the predator. But that drastically *under*estimates the bird's grasp of the situation. Clever experiments on piping plovers by Ristau (1983, 1991) using a remote-controlled toy dune buggy with a stuffed raccoon mounted on it, demonstrated that the plover closely monitors the predator's attention (gaze direction), and modulates its injury feigning, raising the intensity and then letting the predator get closer if the predator shows signs of abandoning the hunt. And, of course, she flies away at the opportune moment, once the predator is some distance from her nest. The bird doesn't need to know the whole rationale, but it does recognize and respond appropriately to some of the conditions alluded to in the rationale. The behavior is neither a simple "knee-jerk" reflex inherited from her ancestors nor a wily scheme figured out in her rational mind; it is an evolution-designed routine with variables that respond to details in the circumstances, details that the sophisticated soliloquy captures—without excess—in the rationale of that design.

The free-floating rationale answers the reverse-engineering question: *Why* is this routine organized like this? If we are squeamish about anthropomorphism, we can pretend to put the answer in somewhat less "mentalistic" terms by liberal use of scare quotes: The routine is an "attention-grabbing" behavior that depends for its success on the likely "goals" and "perceptions" of a predator, designed to provoke the predator into "approaching" the plover and thus distancing itself from the nest; by "monitoring" the predator's "attention" and modulating the behavior to maintain the predator's "interest," the plover typically succeeds in preventing the predation of its young. (This long-winded answer is only superficially more "scientific" than the intentional-stance version expressed in the soliloquy; the two explanations depend on the same distinctions, the same optimality assumptions, and the same informational demands.) Further empirical research may reveal further appropriate sensitivities, or it may reveal the foibles in this cobbled-together device. There is some evidence that piping plovers "know enough" not to engage in injury

feigning when a cow approaches, but instead fly *at* the cow, pushing, not luring, it away from the nest. Would a plover resist the urge to put on an injury-feigning display if it could see that an actually injured bird, or other vulnerable prey item, had already captured the attention of the predator? Or even more wonderful, as suggested by David Haig (2014, personal correspondence):

> One could imagine a bird with an actual broken wing unconvincingly attempting to escape with the intention that the predator interpret its actions as "this is a broken wing display therefore the bird is not easy prey but a nest is near." If the predator started to search for a nest, then the predator would have recognized that the bird's actions were a text but misunderstood the bird's motives. The interpretation of the text is "wrong" for the predator but "right" for the bird. The text has achieved the bird's intention but foiled that of the predator who has been deliberately misled.

Haig speaks unguardedly of the bird's motives and intentions and of the predator's "interpretation" of the "text," recognizing that the task of thinking up these varied opportunities for further experiments and observations *depends on* our adoption of the intentional stance, but also appreciating that there is a graceful and gradual trade-off between interpreting animals (or for that matter, plants or robots or computers) as *themselves* harboring the reasons and the reasoning, and relegating the rationale to Mother Nature, as a free-floating rationale exposed by the mindless design-mining of natural selection.

The *meaning* of the injury-feigning *signal* will have its *intended* effect only if the predator does not *recognize* it to be a *signal* but *interprets* it as an *unintentional* behavior—and this is true whether or not the bird or the predator understands the situation the way we do. It is the *risk* that the predator will *catch on* that creates the selection pressure for better *acting* by the deceptive bird. Similarly, the strikingly realistic "eye-spots" on the wings of butterflies owe their

verisimilitude to the visual acuity of their predators, but of course the butterflies are the clueless beneficiaries of their deceptive gear. The deceptive rationale of the eye-spots is there all the same, and to say it is *there* is to say that there is a domain within which it is *predictive* and, hence, explanatory. (For a related discussion, see Bennett 1976, §§ 52, 53, 62.) We may fail to notice this just because of the obviousness of what we can predict: For example, in an environmental niche with bats but not birds for predators, we don't expect moths with eye-spots (for as any rational deceiver knows, visual sleight of hand is wasted on the blind and myopic).

Comprehension comes in degrees

The time has come to reconsider the slogan *competence without comprehension*. Since cognitive competence is often assumed to be an effect of comprehension, I went out of my way to establish that this familiar assumption is pretty much backward: competence comes first. Comprehension is not the *source* of competence or the *active ingredient* in competence; comprehension is *composed* of competences. We have already considered the possibility of granting a smidgen or two of comprehension to systems that are particularly clever in the ways that they marshal their competences but that may play into the misleading image of comprehension as a separable element or phenomenon kindled somehow by mounting competence.

The idea of comprehension or understanding as a separate, stand-alone, mental marvel is ancient but obsolete. (Think of Descartes's *res cogitans*, or Kant's *Critique of Pure Reason*, or Dilthey's *Verstehen*—which is just the German word for understanding, but since, like all German nouns, it is capitalized, when it is said with furrowed brow, it conjures up in many minds a Bulwark against Reductionism and Positivism, a Humanistic alternative to Science.) The illusion that understanding is some additional, separable mental phenomenon (over and above the set of relevant competences, including the meta-competence to exercise the other competences at appropriate times) is fostered by the *aha!* phenomenon, or eureka

effect—that delightful moment when you suddenly recognize that you *do* understand something that has heretofore baffled you. This psychological phenomenon is perfectly real and has been studied by psychologists for decades. Such an experience of an abrupt onset of understanding can easily be misinterpreted as a demonstration that understanding is a *kind of experience* (as if suddenly learning you were allergic to peanuts would show that allergies are a kind of feeling), and it has led some thinkers to insist that there can be no genuine comprehension without consciousness (Searle [1992] is the most influential). Then, if you were to think that it is obvious that consciousness, whatever it is, sunders the universe in two—everything is either conscious or not conscious; consciousness does not admit of degrees—it would stand to reason that comprehension, *real* comprehension, is enjoyed only by conscious beings. Robots understand nothing, carrots understand nothing, bacteria understand nothing, oysters, well, we don't know *yet*—it all depends on whether oysters are conscious; if not, then their competences, admirable though they are, are competences utterly without comprehension.

I recommend we discard this way of thinking. This well-nigh magical concept of comprehension has no utility, no application in the real world. But the distinction between comprehension and incomprehension is still important, and we can salvage it by the well-tested Darwinian perspective of gradualism: comprehension comes in degrees. At one extreme we have the bacterium's sorta comprehension of the quorum-sensing signals it responds to (Miller and Bassler 2001) and the computer's sorta comprehension of the "ADD" instruction. At the other extreme we have Jane Austen's comprehension of the interplay of personal and social forces in the emotional states of people and Einstein's comprehension of relativity. But even at the highest levels of competence, comprehension is never absolute. There are always ungrasped implications and unrecognized presuppositions in any mind's mastery of a concept or topic. All comprehension is sorta comprehension from some perspective. I once gave a talk at Fermi Lab in Illinois, to a few hundred

of the world's best physicists and confessed that I only sorta understood Einstein's famous formula:

$$E = mc^2$$

I can do the simple algebraic reformulations, and say what each term refers to, and explain (roughly) what is important about this discovery, but I'm sure any wily physicist could easily expose my incomprehension of some aspects of it. (We professors are good at uncovering the mere sorta understanding of our students via examinations.) I then asked how many in the audience understood it. All hands went up, of course, but one person jumped up and shouted "No, no! We theoretical physicists are the ones who understand it; the experimentalists only think they do!" He had a point. Where understanding is concerned, we all depend on something like a division of labor: we count on experts to have deep, "complete" understanding of difficult concepts we rely on every day, only half-comprehendingly. This is, in fact, as we shall see, one of the key contributions of language to our species' intelligence: the capacity to transmit, faithfully, information we only sorta understand!

We human beings are the champion comprehenders on the planet, and when we try to understand other species, we tend to model their comprehension on our experience, imaginatively filling animals' heads with wise reflections as if the animals were strangely shaped people in fur coats. The Beatrix Potter syndrome, as I have called it, is not restricted to children's literature, though I think every culture on earth has folk tales and nursery stories about talking, thinking animals. We do it because, to a first approximation, *it works*. The intentional stance works whether the rationales it adduces are free floating or explicitly represented in the minds of the agents we are predicting. When a son learns from his father how to figure out what their quarry is attending to and how to foil its vigilance, both are treating the animal as a wise fellow thinker in a battle of wits. But the success of the intentional stance does not depend on this being a faithful representation of what is going on

in the animal's mind except to the extent that whatever is going on in the animal's brain has the competence to detect and respond appropriately to the information in the environment.

The intentional stance gives "the specs" for a mind and leaves the implementation for later. This is particularly clear in the case of a chess-playing computer. "Make me a chess program that not only *knows the rules* and *keeps track of all the pieces* but also *notices* opportunities, *recognizes* gambits, *expects* its opponent to make intelligent moves, *values* the pieces soundly, and *looks out for* traps. How you accomplish that is your problem." We adopt the same noncommittal strategy when dealing with a human chess player. In the midst of a chess match we rarely have hunches about—or bother trying to guess—the detailed thinking of our opponent; we expect her to see what's there to be seen, to notice the important implications of whatever changes, and to have good ways of formulating responses to the moves we choose. We idealize everybody's thinking, and even our own access to reasons, blithely attributing phantom bouts of clever reasoning to ourselves after the fact. We tend to see what we chose to do (a chess move, a purchase, parrying a blow) to have been just the right move at the right time, and we have no difficulty explaining to ourselves and others how we figured it out in advance, but when we do this we may often be snatching a free-floating rationale out of thin air and pasting it, retrospectively, into our subjective experience. Asked, "Why did you do that?," the most honest thing to say is often "I don't know; it just came to me," but we often succumb to the temptation to engage in *whig history*, not settling for *how come* but going for a *what for*.[22]

When we turn to the task of modeling the competences out of which comprehension is composed, we can distinguish four grades, schematically characterized by successive applications of the tactic

22 The useful term *Whig history* refers to interpreting history as a story of progress, typically justifying the chain of events leading to the interpreter's privileged vantage point. For applications of the term to adaptationism in evolutionary biology, both favorably and unfavorably, see Cronin (1992) and Griffiths (1995).

known in computer science as "generate and test." In the first, lowest level we find *Darwinian creatures*, with their competences predesigned and fixed, created by the R&D of evolution by natural selection. They are born "knowing" all they will ever "know"; they are gifted but not learners. Each generation generates variations which are then tested against Nature, with the winners copied more often in the next round. Next come the *Skinnerian creatures*, who have, in addition to their hard-wired dispositions, the key disposition to adjust their behavior in reaction to "reinforcement"; they more or less randomly generate new behaviors to test in the world; those that get reinforced (with positive reward or by the removal of an aversive stimulus—pain or hunger, for instance) are more likely to recur in similar circumstances in the future. Those variants born with the unfortunate disposition to mislabel positive and negative stimuli, fleeing the good stuff and going for the bad stuff, soon eliminate themselves, leaving no progeny. This is "operant conditioning" and B. F. Skinner, the arch-behaviorist, noted its echo of Darwinian evolution, with the generation and testing occurring in the individual during its lifetime but requiring no more *comprehension* (mentalism-*fie!*) than natural selection itself. The capacity to improve one's design by operant conditioning is clearly a fitness-enhancing trait under many circumstances, but also risky, since the organism must blindly try out its options in the cruel world (as blindly as evolution does) and may succumb before it learns anything.

Better still is the next grade, the *Popperian creatures*, who extract information about the cruel world and keep it handy, so they can use it to pretest *hypothetical* behaviors offline, letting "their hypotheses die in their stead" as the philosopher of science Karl Popper once put it. Eventually they must act in the real world, but their first choice is not random, having won the generate-and-test competition trial runs in the internal environment model. Finally, there are the *Gregorian creatures*, named in honor of Richard Gregory, the psychologist who emphasized the role of thinking tools in providing thinkers with what he called "potential intelligence." The Gregorian creature's *Umwelt* is well stocked with thinking tools, both abstract

and concrete: arithmetic and democracy and double-blind studies, and microscopes, maps, and computers. A bird in a cage may see as many words every day (on newspaper lining the cage floor) as a human being does, but the words are not thinking tools in the bird's *Umwelt*.

The merely Darwinian creature is "hard-wired," the beneficiary of clever designs it has no need to understand. We can expose its cluelessness by confronting it with novel variations on the conditions it has been designed by evolution to handle: it learns nothing and flounders helplessly. The Skinnerian creature starts out with some "plasticity," some optionality in a repertoire of behaviors that is incompletely designed at birth; it learns by trial-and-error forays in the world and is hard-wired to favor the forays that have "reinforcing" outcomes. It doesn't have to understand why it now prefers these tried-and-true behaviors when it does; it is the beneficiary of this simple design-improvement ratchet, its own portable Darwinian selection process. The Popperian creature looks before it leaps, testing candidates for action against information about the world it has stored in its brain somehow. This looks more like comprehension because the selective process is both information-sensitive and forward-looking, but the Popperian creature need not understand how or why it engages in this pretesting. The "habit" of "creating forward models" of the world and using them to make decisions and modulate behavior is a fine habit to have, whether or not you understand it. Unless you were a remarkably self-reflective child, you "automatically" engaged in Popperian lookahead and reaped some of its benefits long before you noticed you were doing it. Only with the Gregorian creature do we find the deliberate introduction and use of thinking tools, systematic exploration of possible solutions to problems, and attempts at higher-order control of mental searches. Only we human beings are Gregorian creatures, apparently.

Here is where the hot button of human exceptionalism gets pushed, with fierce disagreements between romantics and killjoys (see chapter 1) about how much comprehension is exhibited by which species or which individual animals. The prevailing but

still tentative conclusion these days among researchers in animal intelligence is that the smartest animals are not "just" Skinnerian creatures but Popperian creatures, capable of *figuring out* some of the clever things they have been observed to do. Corvids (crows, ravens, and their close kin), dolphins and other cetaceans, and primates (apes and monkeys) are the most impressive wild animals so far investigated, with dogs, cats, and parrots leading the pet parade. They engage in exploratory behavior, for instance, getting the lay of the land, and often making landmarks to ease the burden on their memories, stocking their heads with handy local information. They need not know that this is the rationale for their behavior, but they benefit from it by reducing uncertainty, extending their powers of anticipation ("look before you leap" is the free-floating maxim of their design), and thereby improving their competences. The fact that they don't understand the grounds of their own understanding is no barrier to calling it understanding, since we humans are often in the same ignorant state about how we manage to figure out novel things, and that is the very hallmark of understanding: the capacity to apply our lessons to new materials, new topics.

Some animals, like us, have something like an inner workshop in which they can engage in do-it-yourself understanding of the prefabricated designs with which they were born. This idea, that the individual organism has a portable design-improvement facility that is more powerful than brute trial-and-error-and-take-your-lumps, is, I submit, the core of our folk understanding of understanding. It doesn't depend on any assumptions about *conscious experience*, although that is a familiar decoration, an ideological amplification, of the basic concept. We are slowly shedding the habit of thinking that way, thanks in part to Freud's championing of unconscious motivations and other psychological states, and thanks also to cognitive science's detailed modeling of unconscious processes of perceptual inference, memory search, language comprehension, and much else. An *unconscious mind* is no longer seen as a "contradiction in terms"; it's the *conscious* minds that apparently raise all the problems. The puzzle today is "what is consciousness *for* (if anything)?" if

unconscious processes are fully competent to perform all the cognitive operations of perception and control.

To summarize, animals, plants, and even microorganisms are equipped with competences that permit them to deal appropriately with the affordances of their environments. There are free-floating rationales for all these competences, but the organisms need not appreciate or comprehend them to benefit from them, nor do they need to be conscious of them. In animals with more complex behaviors, the degree of versatility and variability exhibited can justify attributing a sort of behavioral comprehension to them so long as we don't make the mistake of thinking of comprehension as some sort of stand-alone talent, a source of competence rather than a manifestation of competence.

In part II, we zero in on the evolution of us Gregorian creatures, the reflective users of thinking tools. This development is a giant leap of cognitive competence, putting the human species in a unique niche, but like all evolutionary processes it must be composed of a series of unforeseen and unintended steps, with "full" comprehension a latecomer, not leading the way until very recently.