1

STANDARD EQUIPMENT

Why are there so many robots in fiction, but none in real life? I would pay a lot for a robot that could put away the dishes or run simple errands. But I will not have the opportunity in this century, and probably not in the next one either. There are, of course, robots that weld or spray-paint on assembly lines and that roll through laboratory hallways; my question is about the machines that walk, talk, see, and think, often better than their human masters. Since 1920, when Karel Čapek coined the word *robot* in his play *R.U.R.*, dramatists have freely conjured them up: Speedy, Cutie, and Dave in Isaac Asimov's *I, Robot*, Robbie in *Forbidden Planet*, the flailing canister in *Lost in Space*, the daleks in *Dr. Who*, Rosie the Maid in *The Jetsons*, Nomad in *Star Trek*, Hymie in *Get Smart*, the vacant butlers and bickering haberdashers in *Sleeper*, R2D2 and C3PO in *Star Wars*, the Terminator in *The Terminator*, Lieutenant Commander Data in *Star Trek: The Next Generation*, and the wisecracking film critics in *Mystery Science Theater 3000*.

This book is not about robots; it is about the human mind. I will try to explain what the mind is, where it came from, and how it lets us see, think, feel, interact, and pursue higher callings like art, religion, and philosophy. On the way I will try to throw light on distinctively human quirks. Why do memories fade? How does makeup change the look of a face? Where do ethnic stereotypes come from, and when are they irrational? Why do people lose their tempers? What makes children bratty? Why do fools fall in love? What makes us laugh? And why do people believe in ghosts and spirits?

But the gap between robots in imagination and in reality is my starting point, for it shows the first step we must take in knowing ourselves: appreciating the fantastically complex design behind feats of mental life we take for granted. The reason there are no humanlike robots is not that the very idea of a mechanical mind is misguided. It is that the engineering problems that we humans solve as we see and walk and plan and make it through the day are far more challenging than landing on the moon or sequencing the human genome. Nature, once again, has found ingenious solutions that human engineers cannot yet duplicate. When Hamlet says, "What a piece of work is a man! how noble in reason! how infinite in faculty! in form and moving how express and admirable!" we should direct our awe not at Shakespeare or Mozart or Einstein or Kareem Abdul-Jabbar but at a four-year old carrying out a request to put a toy on a shelf.

In a well-designed system, the components are black boxes that perform their functions as if by magic. That is no less true of the mind. The faculty with which we ponder the world has no ability to peer inside itself or our other faculties to see what makes them tick. That makes us the victims of an illusion: that our own psychology comes from some divine force or mysterious essence or almighty principle. In the Jewish legend of the Golem, a clay figure was animated when it was fed an inscription of the name of God. The archetype is echoed in many robot stories. The statue of Galatea was brought to life by Venus' answer to Pygmalion's prayers; Pinocchio was vivified by the Blue Fairy. Modern versions of the Golem archetype appear in some of the less fanciful stories of science. All of human psychology is said to be explained by a single, omnipotent cause: a large brain, culture, language, socialization, learning, complexity, self-organization, neural-network dynamics.

I want to convince you that our minds are not animated by some godly vapor or single wonder principle. The mind, like the Apollo spacecraft, is designed to solve many engineering problems, and thus is packed with high-tech systems each contrived to overcome its own obstacles. I begin by laying out these problems, which are both design specs for a robot and the subject matter of psychology. For I believe that the discovery by cognitive science and artificial intelligence of the technical challenges overcome by our mundane mental activity is one of the great revelations of science, an awakening of the imagination comparable to learning that the universe is made up of billions of galaxies or that a drop of pond water teems with microscopic life.

THE ROBOT CHALLENGE

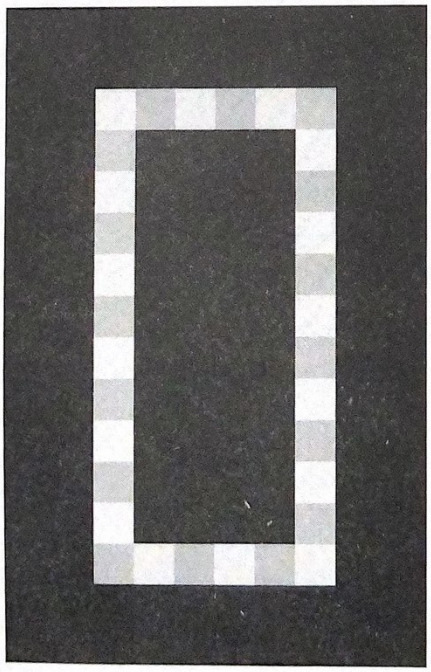
What does it take to build a robot? Let's put aside superhuman abilities like calculating planetary orbits and begin with the simple human ones: seeing, walking, grasping, thinking about objects and people, and planning how to act.

In movies we are often shown a scene from a robot's-eye view, with the help of cinematic conventions like fish-eye distortion or crosshairs. That is fine for us, the audience, who already have functioning eyes and brains. But it is no help to the robot's innards. The robot does not house an audience of little people—homunculi—gazing at the picture and telling the robot what they are seeing. If you could see the world through a robot's eyes, it would look not like a movie picture decorated with crosshairs but something like this:

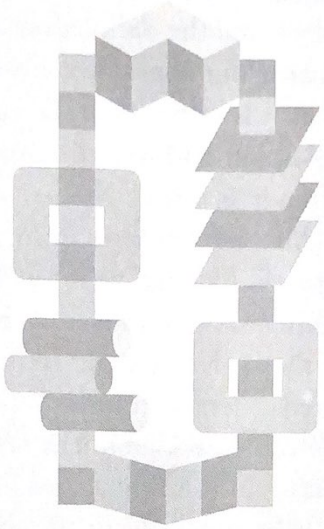
```
225 221 216 219 219 214 207 218 219 220 207 155 136 135
213 206 213 223 208 217 223 221 223 216 195 156 141 130
206 217 210 216 224 223 228 230 234 216 207 157 136 132
211 213 221 223 220 222 237 216 219 220 176 149 137 132
221 229 218 230 228 214 213 209 198 224 161 140 133 127
220 219 224 220 219 215 215 206 206 221 159 143 133 131
221 215 211 214 220 218 221 212 218 204 148 141 131 130
214 211 211 218 214 220 226 216 223 209 143 141 141 124
211 208 223 213 216 226 231 230 241 199 153 141 136 125
200 224 219 215 217 224 232 241 240 211 150 139 128 132
204 206 208 205 233 241 241 252 242 192 151 141 133 130
200 205 201 216 232 248 255 246 231 210 149 141 132 126
191 194 209 238 245 255 249 235 238 197 146 139 130 132
189 199 200 227 239 237 235 236 247 192 145 142 124 133
198 196 209 211 210 215 236 240 232 177 142 137 135 124
198 203 205 208 211 224 226 240 210 160 139 132 129 130
216 209 214 220 210 231 245 219 169 143 148 129 128 136
211 210 217 218 214 227 244 221 162 140 139 129 133 131
215 210 216 216 209 220 248 200 156 139 131 129 139 128
219 220 211 208 205 209 240 217 154 141 127 130 124 142
229 224 212 214 220 229 234 208 151 145 128 128 142 122
252 224 222 224 233 244 228 213 143 141 135 128 131 129
255 235 230 249 253 240 228 193 147 139 132 128 136 125
250 245 238 245 246 235 235 190 139 136 134 135 126 130
240 238 233 232 235 255 246 168 156 144 129 127 136 134
```

Each number represents the brightness of one of the millions of tiny patches making up the visual field. The smaller numbers come from darker patches, the larger numbers from brighter patches. The numbers shown in the array are the actual signals coming from an electronic camera trained on a person's hand, though they could just as well be the firing rates of some of the nerve fibers coming from the eye to the brain as a person looks at a hand. For a robot brain—or a human brain—to recognize objects and not bump into them, it must crunch these numbers and guess what kinds of objects in the world reflected the light that gave rise to them. The problem is humblingly difficult.

First, a visual system must locate where an object ends and the backdrop begins. But the world is not a coloring book, with black outlines around solid regions. The world as it is projected into our eyes is a mosaic of tiny shaded patches. Perhaps, one could guess, the visual brain looks for regions where a quilt of large numbers (a brighter region) abuts a quilt of small numbers (a darker region). You can discern such a boundary in the square of numbers; it runs diagonally from the top right to the bottom center. Most of the time, unfortunately, you would not have found the edge of an object, where it gives way to empty space. The juxtaposition of large and small numbers could have come from many distinct arrangements of matter. This drawing, devised by the psychologists Pawan Sinha and Edward Adelson, appears to show a ring of light gray and dark gray tiles.



In fact, it is a rectangular cutout in a black cover through which you are looking at part of a scene. In the next drawing the cover has been removed, and you can see that each pair of side-by-side gray squares comes from a different arrangement of objects.



Big numbers next to small numbers can come from an object standing in front of another object, dark paper lying on light paper, a surface painted two shades of gray, two objects touching side by side, gray cellophane on a white page, an inside or outside corner where two walls meet, or a shadow. Somehow the brain must solve the chicken-and-egg problem of identifying three-dimensional objects from the patches on the retina *and* determining what each patch is (shadow or paint, crease or overlay, clear or opaque) from knowledge of what object the patch is part of.

The difficulties have just begun. Once we have carved the visual world into objects, we need to know what they are made of, say, snow versus coal. At first glance the problem looks simple. If large numbers come from bright regions and small numbers come from dark regions, then large number equals white equals snow and small number equals black equals coal, right? Wrong. The amount of light hitting a spot on the retina depends not only on how pale or dark the object is but also on how bright or dim the light illuminating the object is. A photographer's light meter would show you that more light bounces off a lump of coal outdoors than off a snowball indoors. That is why people are so often disappointed by their snapshots and why photography is such a complicated craft. The camera does not lie; left to its own devices, it renders outdoor

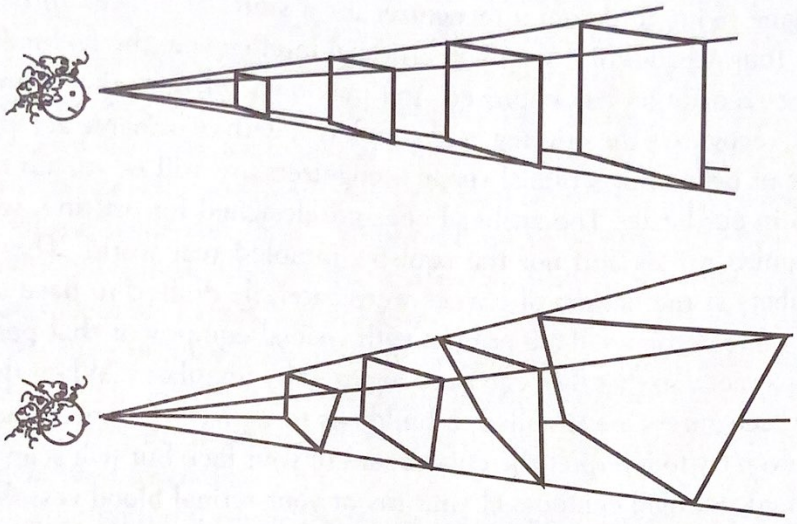
scenes as milk and indoor scenes as mud. Photographers, and sometimes microchips inside the camera, coax a realistic image out of the film with tricks like adjustable shutter timing, lens apertures, film speeds, flashes, and darkroom manipulations.

Our visual system does much better. Somehow it lets us see the bright outdoor coal as black and the dark indoor snowball as white. That is a happy outcome, because our conscious sensation of color and lightness matches the world as it is rather than the world as it presents itself to the eye. The snowball is soft and wet and prone to melt whether it is indoors or out, and we see it as white whether it is indoors or out. The coal is always hard and dirty and prone to burn, and we always see it as black. The harmony between how the world *looks* and how the world *is* must be an achievement of our neural wizardry, because black and white don't simply announce themselves on the retina. In case you are still skeptical, here is an everyday demonstration. When a television set is off, the screen is a pale greenish gray. When it is on, some of the phosphor dots give off light, painting in the bright areas of the picture. But the other dots do not suck light and paint in the dark areas; they just stay gray. The areas that you see as black are in fact just the pale shade of the picture tube when the set was off. The blackness is a figment, a product of the brain circuitry that ordinarily allows you to see coal as coal. Television engineers exploited that circuitry when they designed the screen.

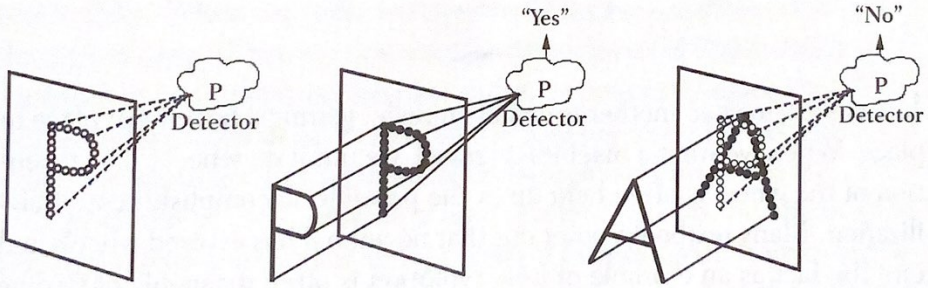
The next problem is seeing in depth. Our eyes squash the three-dimensional world into a pair of two-dimensional retinal images, and the third dimension must be reconstituted by the brain. But there are no telltale signs in the patches on the retina that reveal how far away a surface is. A stamp in your palm can project the same square on your retina as a chair across the room or a building miles away (first drawing, page 9). A cutting board viewed head-on can project the same trapezoid as various irregular shards held at a slant (second drawing, page 9).

You can feel the force of this fact of geometry, and of the neural mechanism that copes with it, by staring at a lightbulb for a few seconds or looking at a camera as the flash goes off, which temporarily bleaches a patch onto your retina. If you now look at the page in front of you, the afterimage adheres to it and appears to be an inch or two across. If you look up at the wall, the afterimage appears several feet long. If you look at the sky, it is the size of a cloud.

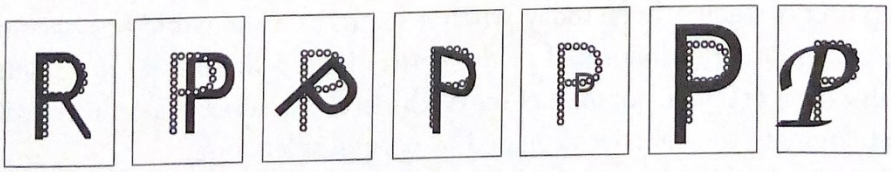
Finally, how might a vision module recognize the objects out there in the world, so that the robot can name them or recall what they do? The



obvious solution is to build a template or cutout for each object that duplicates its shape. When an object appears, its projection on the retina would fit its own template like a round peg in a round hole. The template would be labeled with the name of the shape—in this case, "the letter *P*"—and whenever a shape matches it, the template announces the name:



Alas, this simple device malfunctions in both possible ways. It sees *P*'s that aren't there; for example, it gives a false alarm to the *R* shown in the first square below. And it fails to see *P*'s that are there; for example, it misses the letter when it is shifted, tilted, slanted, too far, too near, or too fancy:



And these problems arise with a nice, crisp letter of the alphabet. Imagine trying to design a recognizer for a shirt, or a face! To be sure, after four decades of research in artificial intelligence, the technology of shape recognition has improved. You may own software that scans in a page, recognizes the printing, and converts it with reasonable accuracy to a file of bytes. But artificial shape recognizers are still no match for the ones in our heads. The artificial ones are designed for pristine, easy-to-recognize worlds and not the squishy, jumbled real world. The funny numbers at the bottom of checks were carefully drafted to have shapes that don't overlap and are printed with special equipment that positions them exactly so that they can be recognized by templates. When the first face recognizers are installed in buildings to replace doormen, they will not even try to interpret the chiaroscuro of your face but will scan in the hard-edged, rigid contours of your iris or your retinal blood vessels. Our brains, in contrast, keep a record of the shape of every face we know (and every letter, animal, tool, and so on), and the record is somehow matched with a retinal image even when the image is distorted in all the ways we have been examining. In Chapter 4 we will explore how the brain accomplishes this magnificent feat.

Let's take a look at another everyday miracle: getting a body from place to place. When we want a machine to move, we put it on wheels. The invention of the wheel is often held up as the proudest accomplishment of civilization. Many textbooks point out that no animal has evolved wheels and cite the fact as an example of how evolution is often incapable of finding the optimal solution to an engineering problem. But it is not a good example at all. Even if nature *could* have evolved a moose on wheels, it surely would have opted not to. Wheels are good only in a world with roads and rails. They bog down in any terrain that is soft, slippery, steep, or uneven. Legs are better. Wheels have to roll along an unbroken supporting ridge, but legs can be placed on a series of separate footholds, an extreme example being a ladder. Legs can also be placed to minimize lurching and to step over obstacles. Even today, when it seems as if the world has become a parking lot, only about half of the earth's land is accessible to vehicles with wheels or tracks, but most of the earth's land is accessible to vehicles with feet: animals, the vehicles designed by natural selection.

But legs come with a high price: the software to control them. A wheel, merely by turning, changes its point of support gradually and can bear weight the whole time. A leg has to change its point of support all at once, and the weight has to be unloaded to do so. The motors controlling a leg have to alternate between keeping the foot on the ground while it bears and propels the load and taking the load off to make the leg free to move. All the while they have to keep the center of gravity of the body within the polygon defined by the feet so the body doesn't topple over. The controllers also must minimize the wasteful up-and-down motion that is the bane of horseback riders. In walking windup toys, these problems are crudely solved by a mechanical linkage that converts a rotating shaft into a stepping motion. But the toys cannot adjust to the terrain by finding the best footholds.

Even if we solved these problems, we would have figured out only how to control a walking insect. With six legs, an insect can always keep one tripod on the ground while it lifts the other tripod. At any instant, it is stable. Even four-legged beasts, when they aren't moving too quickly, can keep a tripod on the ground at all times. But as one engineer has put it, "the upright two-footed locomotion of the human being seems almost a recipe for disaster in itself, and demands a remarkable control to make it practicable." When we walk, we repeatedly tip over and break our fall in the nick of time. When we run, we take off in bursts of flight. These aerobatics allow us to plant our feet on widely or erratically spaced footholds that would not prop us up at rest, and to squeeze along narrow paths and jump over obstacles. But no one has yet figured out how we do it.

Controlling an arm presents a new challenge. Grab the shade of an architect's lamp and move it along a straight diagonal path from near you, low on the left, to far from you, high on the right. Look at the rods and hinges as the lamp moves. Though the shade proceeds along a straight line, each rod swings through a complicated arc, swooping rapidly at times, remaining almost stationary at other times, sometimes reversing from a bending to a straightening motion. Now imagine having to do it in reverse: without looking at the shade, you must choreograph the sequence of twists around each joint that would send the shade along a straight path. The trigonometry is frightfully complicated. But your arm is an architect's lamp, and your brain effortlessly solves the equations every time you point. And if you have ever held an architect's lamp by its clamp, you will appreciate that the problem is even harder than what I have described. The lamp flails under its weight as if it had a mind of its

own; so would your arm if your brain did not compensate for its weight, solving a near-intractable physics problem.

A still more remarkable feat is controlling the hand. Nearly two thousand years ago, the Greek physician Galen pointed out the exquisite natural engineering behind the human hand. It is a single tool that manipulates objects of an astonishing range of sizes, shapes, and weights, from a log to a millet seed. "Man handles them all," Galen noted, "as well as if his hands had been made for the sake of each one of them alone." The hand can be configured into a hook grip (to lift a pail), a scissors grip (to hold a cigarette), a five-jaw chuck (to lift a coaster), a three-jaw chuck (to hold a pencil), a two-jaw pad-to-pad chuck (to thread a needle), a two-jaw pad-to-side chuck (to turn a key), a squeeze grip (to hold a hammer), a disc grip (to open a jar), and a spherical grip (to hold a ball). Each grip needs a precise combination of muscle tensions that mold the hand into the right shape and keep it there as the load tries to bend it back. Think of lifting a milk carton. Too loose a grasp, and you drop it; too tight, and you crush it; and with some gentle rocking, you can even use the tugging on your fingertips as a gauge of how much milk is inside! And I won't even begin to talk about the tongue, a boneless water balloon controlled only by squeezing, which can loosen food from a back tooth or perform the ballet that articulates words like *thrilling* and *sixths*.

"A common man marvels at uncommon things; a wise man marvels at the commonplace." Keeping Confucius' dictum in mind, let's continue to look at commonplace human acts with the fresh eye of a robot designer seeking to duplicate them. Pretend that we have somehow built a robot that can see and move. What will it do with what it sees? How should it decide how to act?

An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects, encountered in the past, to the object at hand.

But whenever one tries to program a set of criteria to capture the members of a category, the category disintegrates. Leaving aside slippery concepts like "beauty" or "dialectical materialism," let's look at a textbook

example of a well-defined one: "bachelor." A bachelor, of course, is simply an adult human male who has never been married. But now imagine that a friend asks you to invite some bachelors to her party. What would happen if you used the definition to decide which of the following people to invite?

> Arthur has been living happily with Alice for the last five years. They have a two-year-old daughter and have never officially married.
>
> Bruce was going to be drafted, so he arranged with his friend Barbara to have a justice of the peace marry them so he would be exempt. They have never lived together. He dates a number of women, and plans to have the marriage annulled as soon as he finds someone he wants to marry.
>
> Charlie is 17 years old. He lives at home with his parents and is in high school.
>
> David is 17 years old. He left home at 13, started a small business, and is now a successful young entrepreneur leading a playboy's lifestyle in his penthouse apartment.
>
> Eli and Edgar are homosexual lovers who have been living together for many years.
>
> Faisal is allowed by the law of his native Abu Dhabi to have three wives. He currently has two and is interested in meeting another potential fiancée.
>
> Father Gregory is the bishop of the Catholic cathedral at Groton upon Thames.

The list, which comes from the computer scientist Terry Winograd, shows that the straightforward definition of "bachelor" does not capture our intuitions about who fits the category.

Knowing who is a bachelor is just common sense, but there's nothing common about common sense. Somehow it must find its way into a human or robot brain. And common sense is not simply an almanac about life that can be dictated by a teacher or downloaded like an enormous database. No database could list all the facts we tacitly know, and no one ever taught them to us. You know that when Irving puts the dog in the car, it is no longer in the yard. When Edna goes to church, her head goes with her. If Doug is in the house, he must have gone in through some opening unless he was born there and never left. If Sheila is alive

at 9 A.M. and is alive at 5 P.M., she was also alive at noon. Zebras in the wild never wear underwear. Opening a jar of a new brand of peanut butter will not vaporize the house. People never shove meat thermometers in their ears. A gerbil is smaller than Mt. Kilimanjaro.

An intelligent system, then, cannot be stuffed with trillions of facts. It must be equipped with a smaller list of core truths and a set of rules to deduce their implications. But the rules of common sense, like the categories of common sense, are frustratingly hard to set down. Even the most straightforward ones fail to capture our everyday reasoning. Mavis lives in Chicago and has a son named Fred, and Millie lives in Chicago and has a son named Fred. But whereas the Chicago that Mavis lives in is the same Chicago that Millie lives in, the Fred who is Mavis' son is not the same Fred who is Millie's son. If there's a bag in your car, and a gallon of milk in the bag, there is a gallon of milk in your car. But if there's a person in your car, and a gallon of blood in a person, it would be strange to conclude that there is a gallon of blood in your car.

Even if you were to craft a set of rules that derived only sensible conclusions, it is no easy matter to use them all to guide behavior intelligently. Clearly a thinker cannot apply just one rule at a time. A match gives light; a saw cuts wood; a locked door is opened with a key. But we laugh at the man who lights a match to peer into a fuel tank, who saws off the limb he is sitting on, or who locks his keys in the car and spends the next hour wondering how to get his family out. A thinker has to compute not just the direct effects of an action but the side effects as well.

But a thinker cannot crank out predictions about *all* the side effects, either. The philosopher Daniel Dennett asks us to imagine a robot designed to fetch a spare battery from a room that also contained a time bomb. Version 1 saw that the battery was on a wagon and that if it pulled the wagon out of the room, the battery would come with it. Unfortunately, the bomb was also on the wagon, and the robot failed to deduce that pulling the wagon out brought the bomb out, too. Version 2 was programmed to consider all the side effects of its actions. It had just finished computing that pulling the wagon would not change the color of the room's walls and was proving that the wheels would turn more revolutions than there are wheels on the wagon, when the bomb went off. Version 3 was programmed to distinguish between relevant implications and irrelevant ones. It sat there cranking out millions of implications and putting all the relevant ones on a list of facts to consider and all the irrelevant ones on a list of facts to ignore, as the bomb ticked away.

An intelligent being has to deduce the implications of what it knows, but only the *relevant* implications. Dennett points out that this requirement poses a deep problem not only for robot design but for epistemology, the analysis of how we know. The problem escaped the notice of generations of philosophers, who were left complacent by the illusory effortlessness of their own common sense. Only when artificial intelligence researchers tried to duplicate common sense in computers, the ultimate blank slate, did the conundrum, now called "the frame problem," come to light. Yet somehow we all solve the frame problem whenever we use our common sense.

Imagine that we have somehow overcome these challenges and have a machine with sight, motor coordination, and common sense. Now we must figure out how the robot will put them to use. We have to give it motives.

What should a robot want? The classic answer is Isaac Asimov's Fundamental Rules of Robotics, "the three rules that are built most deeply into a robot's positronic brain."

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Asimov insightfully noticed that self-preservation, that universal biological imperative, does not automatically emerge in a complex system. It has to be programmed in (in this case, as the Third Law). After all, it is just as easy to build a robot that lets itself go to pot or eliminates a malfunction by committing suicide as it is to build a robot that always looks out for Number One. Perhaps easier; robot-makers sometimes watch in horror as their creations cheerfully shear off limbs or flatten themselves against walls, and a good proportion of the world's most intelligent machines are kamikaze cruise missiles and smart bombs.

But the need for the other two laws is far from obvious. Why give a

robot an order to obey orders—why aren't the original orders enough? Why command a robot not to do harm—wouldn't it be easier never to command it to *do* harm in the first place? Does the universe contain a mysterious force pulling entities toward malevolence, so that a positronic brain must be programmed to withstand it? Do intelligent beings inevitably develop an attitude problem?

In this case Asimov, like generations of thinkers, like all of us, was unable to step outside his own thought processes and see them as artifacts of how our minds were put together rather than as inescapable laws of the universe. Man's capacity for evil is never far from our minds, and it is easy to think that evil just comes along with intelligence as part of its very essence. It is a recurring theme in our cultural tradition: Adam and Eve eating the fruit of the tree of knowledge, Promethean fire and Pandora's box, the rampaging Golem, Faust's bargain, the Sorcerer's Apprentice, the adventures of Pinocchio, Frankenstein's monster, the murderous apes and mutinous HAL of *2001: A Space Odyssey*. From the 1950s through the 1980s, countless films in the computer-runs-amok genre captured a popular fear that the exotic mainframes of the era would get smarter and more powerful and someday turn on us.

Now that computers really *have* become smarter and more powerful, the anxiety has waned. Today's ubiquitous, networked computers have an unprecedented ability to do mischief should they ever go to the bad. But the only mayhem comes from unpredictable chaos or from human malice in the form of viruses. We no longer worry about electronic serial killers or subversive silicon cabals because we are beginning to appreciate that malevolence—like vision, motor coordination, and common sense—does not come free with computation but has to be programmed in. The computer running WordPerfect on your desk will continue to fill paragraphs for as long as it does anything at all. Its software will not insidiously mutate into depravity like the picture of Dorian Gray.

Even if it could, why would it want to? To get—what? More floppy disks? Control over the nation's railroad system? Gratification of a desire to commit senseless violence against laser-printer repairmen? And wouldn't it have to worry about reprisals from technicians who with the turn of a screwdriver could leave it pathetically singing "A Bicycle Built for Two"? A network of computers, perhaps, could discover the safety in numbers and plot an organized takeover—but what would make one computer volunteer to fire the data packet heard round the world and risk early martyrdom? And what would prevent the coalition from being

undermined by silicon draft-dodgers and conscientious objectors? Aggression, like every other part of human behavior we take for granted, is a challenging engineering problem!

But then, so are the kinder, gentler motives. How would you design a robot to obey Asimov's injunction never to allow a human being to come to harm through inaction? Michael Frayn's 1965 novel *The Tin Men* is set in a robotics laboratory, and the engineers in the Ethics Wing, Macintosh, Goldwasser, and Sinson, are testing the altruism of their robots. They have taken a bit too literally the hypothetical dilemma in every moral philosophy textbook in which two people are in a lifeboat built for one and both will die unless one bails out. So they place each robot in a raft with another occupant, lower the raft into a tank, and observe what happens.

> [The] first attempt, Samaritan I, had pushed itself overboard with great alacrity, but it had gone overboard to save anything which happened to be next to it on the raft, from seven stone of lima beans to twelve stone of wet seaweed. After many weeks of stubborn argument Macintosh had conceded that the lack of discrimination was unsatisfactory, and he had abandoned Samaritan I and developed Samaritan II, which would sacrifice itself only for an organism at least as complicated as itself.
>
> The raft stopped, revolving slowly, a few inches above the water. "Drop it," cried Macintosh.
>
> The raft hit the water with a sharp report. Sinson and Samaritan sat perfectly still. Gradually the raft settled in the water, until a thin tide began to wash over the top of it. At once Samaritan leaned forward and seized Sinson's head. In four neat movements it measured the size of his skull, then paused, computing. Then, with a decisive click, it rolled sideways off the raft and sank without hesitation to the bottom of the tank.

But as the Samaritan II robots came to behave like the moral agents in the philosophy books, it became less and less clear that they were really moral at all. Macintosh explained why he did not simply tie a rope around the self-sacrificing robot to make it easier to retrieve: "I don't want it to know that it's going to be saved. It would invalidate its decision to sacrifice itself. . . . So, every now and then I leave one of them in instead of fishing it out. To show the others I mean business. I've written off two this week." Working out what it would take to program goodness into a robot shows not only how much machinery it takes to be good but how slippery the concept of goodness is to start with.

And what about the most caring motive of all? The weak-willed com-

puters of 1960s pop culture were not tempted only by selfishness and power, as we see in the comedian Allan Sherman's song "Automation," sung to the tune of "Fascination":

> It was automation, I know.
> That was what was making the factory go.
> It was IBM, it was Univac,
> It was all those gears going clickety clack, dear.
> I thought automation was keen
> Till you were replaced by a ten-ton machine.
> It was a computer that tore us apart, dear,
> Automation broke my heart. . . .
>
> It was automation, I'm told,
> That's why I got fired and I'm out in the cold.
> How could I have known, when the 503
> Started in to blink, it was winking at me, dear?
> I thought it was just some mishap
> When it sidled over and sat on my lap.
> But when it said "I love you" and gave me a hug, dear,
> That's when I pulled out . . . its . . . plug.

But for all its moonstruck madness, love is no bug or crash or malfunction. The mind is never so wonderfully concentrated as when it turns to love, and there must be intricate calculations that carry out the peculiar logic of attraction, infatuation, courtship, coyness, surrender, commitment, malaise, philandering, jealousy, desertion, and heartbreak. And in the end, as my grandmother used to say, every pot finds a cover; most people—including, significantly, all of our ancestors—manage to pair up long enough to produce viable children. Imagine how many lines of programming it would take to duplicate that!

Robot design is a kind of consciousness-raising. We tend to be blasé about our mental lives. We open our eyes, and familiar articles present themselves; we will our limbs to move, and objects and bodies float into place; we awaken from a dream, and return to a comfortingly predictable

world; Cupid draws back his bow, and lets his arrow go. But think of what it takes for a hunk of matter to accomplish these improbable outcomes, and you begin to see through the illusion. Sight and action and common sense and violence and morality and love are no accident, no inextricable ingredients of an intelligent essence, no inevitability of information processing. Each is a tour de force, wrought by a high level of targeted design. Hidden behind the panels of consciousness must lie fantastically complex machinery—optical analyzers, motion guidance systems, simulations of the world, databases on people and things, goal-schedulers, conflict-resolvers, and many others. Any explanation of how the mind works that alludes hopefully to some single master force or mind-bestowing elixir like "culture," "learning," or "self-organization" begins to sound hollow, just not up to the demands of the pitiless universe we negotiate so successfully.

The robot challenge hints at a mind loaded with original equipment, but it still may strike you as an argument from the armchair. Do we actually find signs of this intricacy when we look directly at the machinery of the mind and at the blueprints for assembling it? I believe we do, and what we see is as mind-expanding as the robot challenge itself.

When the visual areas of the brain are damaged, for example, the visual world is not simply blurred or riddled with holes. Selected aspects of visual experience are removed while others are left intact. Some patients see a complete world but pay attention only to half of it. They eat food from the right side of the plate, shave only the right cheek, and draw a clock with twelve digits squished into the right half. Other patients lose their sensation of color, but they do not see the world as an arty black-and-white movie. Surfaces look grimy and rat-colored to them, killing their appetite and their libido. Still others can see objects change their positions but cannot see them move—a syndrome that a philosopher once tried to convince me was logically impossible! The stream from a teapot does not flow but looks like an icicle; the cup does not gradually fill with tea but is empty and then suddenly full.

Other patients cannot recognize the objects they see: their world is like handwriting they cannot decipher. They copy a bird faithfully but identify it as a tree stump. A cigarette lighter is a mystery until it is lit. When they try to weed the garden, they pull out the roses. Some patients can recognize inanimate objects but cannot recognize faces. The patient deduces that the visage in the mirror must be his, but does not viscerally recognize himself. He identifies John F. Kennedy as Martin Luther King,

and asks his wife to wear a ribbon at a party so he can find her when it is time to leave. Stranger still is the patient who recognizes the face but not the person: he sees his wife as an amazingly convincing impostor.

These syndromes are caused by an injury, usually a stroke, to one or more of the thirty brain areas that compose the primate visual system. Some areas specialize in color and form, others in where an object is, others in what an object is, still others in how it moves. A seeing robot cannot be built with just the fish-eye viewfinder of the movies, and it is no surprise to discover that humans were not built that way either. When we gaze at the world, we do not fathom the many layers of apparatus that underlie our unified visual experience, until neurological disease dissects them for us.

Another expansion of our vista comes from the startling similarities between identical twins, who share the genetic recipes that build the mind. Their minds are astonishingly alike, and not just in gross measures like IQ and personality traits like neuroticism and introversion. They are alike in talents such as spelling and mathematics, in opinions on questions such as apartheid, the death penalty, and working mothers, and in their career choices, hobbies, vices, religious commitments, and tastes in dating. Identical twins are far more alike than fraternal twins, who share only half their genetic recipes, and most strikingly, they are almost as alike when they are reared apart as when they are reared together. Identical twins separated at birth share traits like entering the water backwards and only up to their knees, sitting out elections because they feel insufficiently informed, obsessively counting everything in sight, becoming captain of the volunteer fire department, and leaving little love notes around the house for their wives.

People find these discoveries arresting, even incredible. The discoveries cast doubt on the autonomous "I" that we all feel hovering above our bodies, making choices as we proceed through life and affected only by our past and present environments. Surely the mind does not come equipped with so many small parts that it could predestine us to flush the toilet before and after using it or to sneeze playfully in crowded elevators, to take two other traits shared by identical twins reared apart. But apparently it does. The far-reaching effects of the genes have been documented in scores of studies and show up no matter how one tests for them: by comparing twins reared apart and reared together, by comparing identical and fraternal twins, or by comparing adopted and biological children. And despite what critics sometimes claim, the effects are not

products of coincidence, fraud, or subtle similarities in the family environments (such as adoption agencies striving to place identical twins in homes that both encourage walking into the ocean backwards). The findings, of course, can be misinterpreted in many ways, such as by imagining a gene for leaving little love notes around the house or by concluding that people are unaffected by their experiences. And because this research can measure only the ways in which people *differ*, it says little about the design of the mind that all normal people share. But by showing how many ways the mind can vary in its innate structure, the discoveries open our eyes to how much structure the mind must have.

REVERSE-ENGINEERING THE PSYCHE

The complex structure of the mind is the subject of this book. Its key idea can be captured in a sentence: The mind is a system of organs of computation, designed by natural selection to solve the kinds of problems our ancestors faced in their foraging way of life, in particular, understanding and outmaneuvering objects, animals, plants, and other people. The summary can be unpacked into several claims. The mind is what the brain does; specifically, the brain processes information, and thinking is a kind of computation. The mind is organized into modules or mental organs, each with a specialized design that makes it an expert in one arena of interaction with the world. The modules' basic logic is specified by our genetic program. Their operation was shaped by natural selection to solve the problems of the hunting and gathering life led by our ancestors in most of our evolutionary history. The various problems for our ancestors were subtasks of one big problem for their genes, maximizing the number of copies that made it into the next generation.

On this view, psychology is engineering in reverse. In forward-engineering, one designs a machine to do something; in reverse-engineering, one figures out what a machine was designed to do. Reverse-engineering is what the boffins at Sony do when a new product is announced by Panasonic, or vice versa. They buy one, bring it back to the lab, take a screwdriver to it, and try to figure out what all the parts are for and how they combine to make the device work. We all engage in reverse-engineering when we face an interesting new gadget. In rummaging through