KINDS OF MINDS

Toward an Understanding of Consciousness

DANIEL C. DENNETT

BasicBooks
A Division of HarperCollins*Publishers*

The Science Masters Series is a global publishing venture consisting of original science books written by leading scientists and published by a worldwide team of twenty-six publishers assembled by John Brockman. The series was conceived by Anthony Cheetham of Orion Publishers and John Brockman of Brockman Inc., a New York literary agency, and developed in coordination with BasicBooks.

••••••••

The Science Masters name and marks are owned by and licensed to the publisher by Brockman Inc.

••••••••

••••••••

Published by BasicBooks
A Division of HarperCollinsPublishers, Inc.

••••••••

••••••••

Designed by Joan Greenfield

••••••••

FIRST EDITION

••••••••

ISBN 0-465-07350-6

••••••••

96 97 98 99 00 ❖/HC 10 9 8 7 6 5 4 3 2 1

CONTENTS

CHAPTER 1

WHAT KINDS OF MINDS ARE THERE?

KNOWING YOUR OWN MIND

Can we ever really know what is going on in someone else's mind? Can a woman ever know what it is like to be a man? What experiences does a baby have during childbirth? What experiences, if any, does a fetus have in its mother's womb? And what of nonhuman minds? What do horses think about? Why aren't vultures nauseated by the rotting carcasses they eat? When a fish has a hook sticking through its lip, does it hurt the fish as much as it would hurt you, if you had a hook sticking through your lip? Can spiders think, or are they just tiny robots, mindlessly making their elegant webs? For that matter, why couldn't a robot—if it was fancy enough—be conscious? There are robots that can move around and manipulate things almost as adeptly as spiders; could a more complicated robot feel pain, and worry about its future, the way a person can? Or is there some unbridgeable chasm separating the robots (and maybe the spiders and insects and other "clever" but mindless creatures) from those animals that have minds? Could it be that all animals except human beings are really mindless robots? René Descartes notoriously

maintained this in the seventeenth century. Might he have been dead wrong? Could it be that all animals, and even plants—and even bacteria—have minds?

Or, to swing to the other extreme, are we so sure that all human beings have minds? Maybe (to take the most extreme case of all) you're the only mind in the universe; maybe everything else, including the apparent author of this book, is a mere mindless machine. This strange idea first occurred to me when I was a young child, and perhaps it did to you as well. Roughly a third of my students claim that they, too, invented it on their own and mulled it over when they were children. They are often amused to learn that it's such a common philosophical hypothesis that it has a name—*solipsism* (from Latin for "myself alone"). Nobody ever takes solipsism seriously for long, as far as we know, but it does raise an important challenge: *if* we know that solipsism is silly—*if* we know that there are other minds—how do we know?

What kinds of minds are there? And how do we know? The first question is about what exists—about *ontology*, in philosophical parlance; the second question is about our knowledge—about *epistemology*. The goal of this book is not to answer these two questions once and for all, but rather to show why these questions have to be answered together. Philosophers often warn against confusing ontological questions with epistemological questions. What exists is one thing, they say, and what we can know about it is something else. There may be things that are completely unknowable to us, so we must be careful not to treat the limits of our knowledge as sure guides to the limits of what there is. I agree that this is good general advice, but I will argue that we already know enough about minds to know that one of the things that makes them different from everything else in the universe is the *way* we know about them. For instance, you know you have a mind and you know you have a brain, but

these are different kinds of knowledge. You know you have a brain the way you know you have a spleen: by hearsay. You've never seen your spleen or your brain (I would bet), but since the textbooks tell you that all normal human beings have one of each, you conclude that you almost certainly have one of each as well. You are more intimately acquainted with your mind—so intimately that you might even say that you *are* your mind. (That's what Descartes said: he said he was a mind, a *res cogitans,* or thinking thing.) A book or a teacher might tell you what a mind is, but you wouldn't have to take anybody's word for the claim that you had one. If it occurred to you to wonder whether you were normal and had a mind as other people do, you would immediately realize, as Descartes pointed out, that your very wondering this wonder demonstrated beyond all doubt that you did indeed have a mind.

This suggests that each of us knows exactly one mind from the inside, and no two of us know the same mind from the inside. No other kind of thing is known about in that way. And yet this whole discussion so far has been conducted in terms of how *we* know—you and I. It presupposes that solipsism is false. The more we—we—reflect on this presupposition, the more unavoidable it appears. There couldn't be just one mind—or at least not just one mind like *our* minds.

WE MIND-HAVERS, WE MINDERS

If we want to consider the question of whether nonhuman animals have minds, we have to start by asking whether they have minds in some regards like ours, since these are the only minds we know anything about—at this point. (Try asking yourself whether nonhuman animals have flurbs. You

can't even know what the question is, if you don't know what a flurb is supposed to be. Whatever else a mind is, it is supposed to be something like our minds; otherwise we wouldn't call it a mind.) So our minds, the only minds *we* know from the outset, are the standard with which we must begin. Without this agreement, we'll just be fooling ourselves, talking rubbish without knowing it.

When *I* address *you*, I include us both in the class of mind-havers. This unavoidable starting point creates, or acknowledges, an in-group, a class of privileged characters, set off against everything else in the universe. This is almost too obvious to notice, so deeply enshrined is it in our thinking and talking, but I must dwell on it. When there's a *we*, you are not alone; solipsism is false; there's company present. This comes out particularly clearly if we consider some curious variations:

> "We left Houston at dawn, headin' down the road—just me and my truck."

Strange. If this fellow thinks his truck is such a worthy companion that it deserves shelter under the umbrella of "we," he must be very lonely. Either that, or his truck must have been customized in ways that would be the envy of roboticists everywhere. In contrast, "we—just me and my dog" doesn't startle us at all, but "we—just me and my oyster" is hard to take seriously. In other words, we're pretty sure that dogs have minds, and we're dubious that oysters do.

Membership in the class of things that have minds provides an all-important guarantee: the guarantee of a certain sort of moral standing. Only mind-havers can care; only mind-havers can mind what happens. If I do something to you that you don't want me to do, this has moral significance. It matters, because it matters to you. It may not matter much, or your interests may be overridden for all sorts of

reasons, or (if I'm punishing you justly for a misdeed of yours) the fact that you care may actually count *in favor* of my deed. In any event, your caring automatically counts for something in the moral equation. If flowers have minds, then what we do to flowers can matter *to them*, and not just to those who care about what happens to flowers. If nobody cares, then it doesn't matter what happens to flowers.

There are some who would disagree; they would insist that the flowers had some moral standing even if nothing with a mind knew of or cared about their existence. Their beauty, for instance, no matter how unappreciated, is a good thing in itself, and hence should not be destroyed, other things being equal. This is not the view that the beauty of these flowers matters *to God*, for instance, or that it *might* matter to some being whose presence is undetectable by us. It is the view that the beauty matters, *even though it matters to no one*—not to the flowers themselves and not to God or anybody else. I remain unpersuaded, but rather than dismiss this view outright I will note that it is controversial and not widely shared. In contrast, it takes no special pleading at all to get most people to agree that something with a mind has interests that matter. That's why people are so concerned, morally, about the question of what has a mind: any proposed adjustment in the boundary of the class of mind-havers has major ethical significance.

We might make mistakes. We might endow mindless things with minds, or we might ignore a mindful thing in our midst. These mistakes would not be equal. To overattribute minds—to "make friends with" your houseplants or lie awake at night worrying about the welfare of the computer asleep on your desk—is, at worst, a silly error of credulity. To underattribute minds—to disregard or discount or deny the experience, the suffering and joy, the thwarted ambitions and frustrated desires of a mind-having person or animal—would be a terrible sin. After all, how would *you*

feel if you were treated as an inanimate object? (Notice how this rhetorical question appeals to *our* shared status as mind-havers.)

In fact, both errors could have serious moral consequences. If we overattributed minds (if, for instance, we got it into our heads that since bacteria had minds, we couldn't justify killing them), this might lead us to sacrifice the interests of many legitimate interest-holders—our friends, our pets, ourselves—for nothing of genuine moral importance. The abortion debate hinges on just such a quandary; some think it's obvious that a ten-week-old fetus has a mind, and others think it's obvious that it does not. If it does not, then the path is open to argue that it has no more interests than, say, a gangrenous leg or an abscessed tooth—it can be destroyed to save the life (or just to suit the interests) of the mind-haver of which it is a part. If it does already have a mind, then, whatever we decide, we obviously have to consider *its* interests along with the interests of its temporary host. In between these extreme positions lies the real quandary: the fetus will soon develop a mind if left undisturbed, so when do we start counting its *prospective* interests? The relevance of mind-having to the question of moral standing is especially clear in these cases, since if the fetus in question is known to be anencephalic (lacking a brain), this dramatically changes the issue for most people. Not for all. (I am not attempting to settle these moral issues here, but just to show how a common moral opinion amplifies our interest in these questions way beyond normal curiosity.)

The dictates of morality and scientific method pull in opposite directions here. The ethical course is to err on the side of overattribution, just to be safe. The scientific course is to put the burden of proof on the attribution. As a scientist, you can't just *declare*, for instance, that the presence of glutamate molecules (a basic neurotransmitter involved in

signaling between nerve cells) amounts to the presence of mind; you have to prove it, against a background in which the "null hypothesis" is that mind is not present. (*Innocent until proven guilty* is the null hypothesis in our criminal law.) There is substantial disagreement among scientists about which species have what sorts of mind, but even those scientists who are the most ardent champions of consciousness in animals accept this burden of proof—and think they can meet it, by devising and confirming theories that show which animals are conscious. But no such theories are yet confirmed, and in the meantime we can appreciate the discomfort of those who see this agnostic, wait-and-see policy as jeopardizing the moral status of creatures they are *sure* are conscious.

Suppose the question before us were not about the minds of pigeons or bats but about the minds of left-handed people or people with red hair. We would be deeply offended to be told that it had yet to be proved that this category of living thing had the wherewithal for entry into the privileged class of mind-havers. Many people are similarly outraged by the demand for proof of mind-having in nonhuman species, but if they're honest with themselves they will grant that they, too, see the need for such proof in the case of, say, jellyfish or amoebas or daisies; so we agree on the principle, and they're just taking umbrage at its application to creatures so very much like us. We can allay their misgivings somewhat by agreeing that we should err well on the side of inclusiveness in all our policies, until the facts are in; still, the price you must pay for scientific confirmation of your favorite hypothesis about animal minds is the risk of scientific disconfirmation.

WORDS AND MINDS

It is beyond serious dispute, however, that you and I each have a mind. How do I know you have a mind? Because anybody who can understand my words is automatically addressed by my pronoun "you," and only things with minds can understand. There are computer-driven devices that can read books for the blind: they convert a page of visible text into a stream of audible words, but they don't understand the words they read and hence are not addressed by any "you" they encounter; it passes right through them and addresses whoever listens to—and understands—the stream of spoken words. That's how I know that you, gentle reader/listener, have a mind. So do I. Take my word for it.

In fact that's what we routinely do: we take each other's words as settling beyond any reasonable doubt the question of whether we each have minds. Why should words be so convincing? Because they are such powerful resolvers of doubts and ambiguities. You see somebody coming toward you, scowling and waving an ax. You wonder, What's his problem? Is he going to attack me? Is he mistaking me for somebody else? Ask him. Perhaps he will confirm your worst fears, or perhaps he will tell you he has given up trying to unlock his car (which you're standing in front of) and has returned with his ax to break the window. You may not believe him when he says it's his car, not somebody else's, but further conversation—if you decide not to run away—is bound to resolve your doubts and clarify the situation in ways that would be all but impossible if you and he were unable to communicate verbally. Suppose you try asking him, but it turns out that he doesn't speak your language. Perhaps you will then both resort to gestures and miming. These techniques, used with ingenuity, will take you far, but they're a poor substitute for language—just reflect on how

eagerly you would both seek to confirm your hard-won understanding if a bilingual interpreter were to come along. A few relayed questions and answers would not just allay any residual uncertainty but would add details that could not be conveyed in any other way: "When he saw you put one hand on your chest and push out with your other hand, he thought you meant that you were ill; he was trying to ask if you wanted him to take you to a doctor once he'd broken the window and retrieved his keys. That business with his fingers in his ears was his attempt to convey a stethoscope." Ah, it all falls into place now, thanks to a few words.

People often emphasize the difficulty of accurate and reliable translation between human languages. Human cultures, we are told, are too different, too "incommensurable," to permit the meanings available to one speaker to be perfectly shared with another. No doubt translation always falls somewhat short of perfection, but this may not matter much in the larger scheme of things. Perfect translation may be impossible, but good translation is achieved every day—routinely, in fact. Good translation can be objectively distinguished from not-so-good translation and from bad translation, and it permits all human beings, regardless of race, culture, age, gender, or experience, to unite more closely with one another than individuals of any other species can. We human beings share a subjective world—and know that we do—in a way that is entirely beyond the capacities of any other creatures on the planet, because we can talk to one another. Human beings who don't (yet) have a language in which to communicate are the exception, and that's why we have a particular problem figuring out what it's like to be a newborn baby or a deaf-mute.

Conversation unites us. We can all know a great deal about what it's like to be a Norwegian fisherman or a Nigerian taxi driver, an eighty-year-old nun or a five-year-old boy blind from birth, a chess master or a prostitute or a fighter

pilot. We can know much more about these topics than we can know about what it's like (if anything) to be a dolphin, a bat, or even a chimpanzee. No matter how different from one another we people are, scattered around the globe, we can explore our differences and communicate about them. No matter how similar to one another wildebeests are, standing shoulder to shoulder in a herd, they cannot know much of anything about their similarities, let alone their differences. They cannot compare notes. They can have similar experiences, side by side, but they really cannot share experiences the way we do.

Some of you may doubt this. Can't animals "instinctively" understand each other in ways we human beings cannot fathom? Certainly some authors have said so. Consider, for instance, Elizabeth Marshall Thomas, who imagines, in *The Hidden Life of Dogs* (1993), that dogs enjoy a wise understanding of their own ways. One example: "For reasons known to dogs but not to us, many dog mothers won't mate with their sons." (p. 76). Their instinctive resistance to such inbreeding is not in doubt, but what gives her the idea that dogs have any more insight into the reasons for their instincts than we have into ours? There are many things we feel strongly and instinctively disinclined to do, with no inkling about why we feel that way. To suppose without proof that dogs have more insight into their urges than we do is to ignore the null hypothesis in an unacceptable way—if we are asking a scientific question. As we shall see, very simple organisms may be attuned to their environments and to each other in strikingly apt ways without having the slightest appreciation of their attunement. We *already* know from conversation, however, that people are typically capable of a very high order of understanding of themselves and others.

Of course, we can be fooled. People often emphasize the

difficulty of determining whether a speaker is sincere. Words, by being the most powerful tools of communication, are also the most powerful tools of deception and manipulation. But while it may be easy to lie, it's almost as easy to catch a liar—especially when the lies get large and the logistical problem of maintaining the structure of falsehood overwhelms the liar. In fantasy, we can conjure up infinitely powerful deceivers, but the deceptions that are "possible in principle" to such an evil demon can be safely ignored in the real world. It would be just too difficult to make up that much falsehood and maintain it consistently. We *know* that people the world over have much the same likes and dislikes, hopes and fears. We know that they enjoy recollecting favorite events in their lives. We know that they all have rich episodes of waking fantasy, in which they rearrange and revise the details deliberately. We know that they have obsessions, nightmares, and hallucinations. We know that they can be reminded by an aroma or a melody of a specific event in their lives, and that they often talk to themselves silently, without moving their lips. Long before there was scientific psychology, long before there was meticulous observation of and experimentation on human subjects, this was all common knowledge. We have known these facts about people since ancient times, because we have talked it over with them, at great length. We know nothing comparable about the mental lives of any other species, because we can't talk it over with them. We may think we know, but it takes scientific investigation to confirm or refute our traditional hunches.

CHAPTER 2

INTENTIONALITY: THE INTENTIONAL SYSTEMS APPROACH

> I notice something and seek a reason for it: this means originally: I seek an intention in it, and above all someone who has intentions, a subject, a doer: every event a deed—formerly one saw intentions in all events, this is our oldest habit. Do animals also possess it?
>
> Friedrich Nietzsche, *The Will to Power*

SIMPLE BEGINNINGS: THE BIRTH OF AGENCY*

No grain of sand has a mind; a grain of sand is too simple. Even simpler, no carbon atom or water molecule has a mind. I expect no serious disagreement about that. But what about larger molecules? A virus is a single huge molecule, a macromolecule composed of hundreds of thousands or even millions of parts, depending on how small the parts are that we

*Portions of this section are drawn from my 1995 book, *Darwin's Dangerous Idea*, with revisions.

count. These atomic-level parts interact, in their obviously mindless ways, to produce some quite striking effects. Chief among these effects, from the point of view of our investigation, is *self-replication*. Some macromolecules have the amazing ability, if left floating in a suitably well-furnished medium, to mindlessly construct and then shed exact—or nearly exact—copies of themselves. DNA and its ancestor, RNA, are such macromolecules; they are the foundation of all life on this planet and hence a historical precondition for all minds—at least, all minds on this planet. For about a billion years before simple single-celled organisms appeared on earth, there were self-replicating macromolecules, ceaselessly mutating, growing, even repairing themselves, and getting better and better at it—and replicating over and over again.

This is a stupendous feat, still well beyond the capacity of any existing robot. Does that mean that such macromolecules have minds like ours? Certainly not. They're not even alive—they're just huge crystals, from the point of view of chemistry. These gigantic molecules are tiny machines—*macro*molecular *nano*technology. They are, in effect, natural robots. The possibility in principle of a self-replicating robot was mathematically demonstrated by John von Neumann, one of the inventors of the computer, whose brilliant design for a nonliving self-replicator anticipated many of the details of design and construction of RNA and DNA.

Through the microscope of molecular biology, we get to witness the birth of *agency*, in the first macromolecules that have enough complexity to *perform actions*, instead of just lying there *having effects*. Their agency is not fully fledged agency like ours. They know not what they do. We, in contrast, often know full well what we do. At our best—and at our worst—we human agents can perform *intentional* actions, after having deliberated consciously about the reasons for and against. Macromolecular agency is different;

there are reasons for what macromolecules do, but the macromolecules are unaware of those reasons. Their sort of agency is nevertheless the only possible ground from which the seeds of our kind of agency could grow.

There is something alien and vaguely repellent about the quasi agency we discover at this level—all that purposive hustle and bustle, and yet "there's nobody home." The molecular machines perform their amazing stunts, obviously exquisitely designed and just as obviously none the wiser about what they are doing. Consider this account of the activity of an RNA phage—a replicating virus and a modern-day descendant of the earliest self-replicating macromolecules:

> First of all, the virus needs a material in which to pack and protect its own genetic information. Secondly, it needs a means of introducing its information into the host cell. Thirdly, it requires a mechanism for the specific replication of its information in the presence of a vast excess of host cell RNA. Finally, it must arrange for the proliferation of its information, a process that usually leads to the destruction of the host cell. . . . The virus even gets the cell to carry out its replication; its only contribution is one protein factor, specially adapted for the viral RNA. This enzyme does not become active until a "password" on the viral RNA is shown. When it sees this, it reproduces the viral RNA with great efficiency, while ignoring the very much greater number of RNA molecules of the host cell. Consequently the cell is soon flooded with viral RNA. This is packed into the virus' coat protein, which is also synthesized in large quantities, and finally the cell bursts and releases a multitude of progeny virus particles. All this is a programme that runs automatically and is rehearsed down to the smallest detail. (Eigen, 1992, p. 40)

The author, the molecular biologist Manfred Eigen, has helped himself to a rich vocabulary of agency words: in order to reproduce, the virus must "arrange for" the proliferation of its information, and in furthering this goal it creates an enzyme that "sees" its password and "ignores" other molecules. This is poetic license, to be sure; these words have had their meanings stretched for the occasion. But what an irresistible stretch! The agency words draw attention to the most striking features of the phenomena: these macromolecules are *systematic*. Their control systems are not just efficient at what they do; they are appropriately sensitive to variation, opportunistic, ingenious, devious. They can be "fooled," but only by novelties not regularly encountered by their ancestors.

These impersonal, unreflective, robotic, mindless little scraps of molecular machinery are the ultimate basis of all the agency, and hence meaning, and hence consciousness, in the world. It is rare for such a solid and uncontroversial scientific fact to have such potent implications for structuring all subsequent debate about something as controversial and mysterious as minds, so let's pause to remind ourselves of these implications.

There is no longer any serious informed doubt about this: *we are the direct descendants of these self-replicating robots*. We are mammals, and all mammals have descended from reptilian ancestors whose ancestors were fish whose ancestors were marine creatures rather like worms, who descended in turn from simpler multicelled creatures several hundred million years ago, who descended from single-celled creatures who descended from self-replicating macromolecules, about three billion years ago. There is just one family tree, on which all living things that have ever lived on this planet can be found—not just animals, but plants and algae and bacteria as well. You share a common ancestor with every chimpanzee, every worm, every blade of grass,

every redwood tree. Among our progenitors, then, were macromolecules.

To put it vividly, your great-great-. . . grandmother *was* a robot! Not only are you descended from such macromolecular robots but you are composed of them: your hemoglobin molecules, your antibodies, your neurons, your vestibular-ocular reflex machinery—at every level of analysis from the molecular on up, your body (including your brain, of course) is found to be composed of machinery that dumbly does a wonderful, elegantly designed job.

We have ceased to shudder, perhaps, at the scientific vision of viruses and bacteria busily and mindlessly executing their subversive projects—horrid little automata doing their evil deeds. But we should not think that we can take comfort in the thought that *they* are alien invaders, so unlike the more congenial tissues that make up *us*. We are made of the same sorts of automata that invade us—no special halos of humanity distinguish your antibodies from the antigens they combat; your antibodies simply belong to the club that is you, so they fight on your behalf. The billions of neurons that band together to make your brain are cells, the same sort of biological entity as the germs that cause infections, or the yeast cells that multiply in the vat when beer is fermenting or in the dough when bread rises.

Each cell—a tiny agent that can perform a limited number of tasks—is about as mindless as a virus. Can it be that if enough of these dumb homunculi—little men—are put together the result will be a real, conscious person, with a genuine mind? According to modern science, there is no other way of making a real person. Now, it certainly does not follow from the fact that we are descended from robots that we are robots ourselves. After all, we are also direct descendants of fish, and we are not fish; we are direct descendants of bacteria, and we are not bacteria. But unless there is some secret extra ingredient in us (which is what dualists and

vitalists used to think), we are *made of* robots—or, what comes to the same thing, we are each a collection of trillions of macromolecular machines. And all of these are ultimately descended from the original self-replicating macromolecules. So something made of robots *can* exhibit genuine consciousness, because you do if anything does.

To some people, all this seems shocking and unlikely, I realize, but I suspect that they haven't noticed how desperate the alternatives are. Dualism (the view that minds are composed of some nonphysical and utterly mysterious stuff) and vitalism (the view that living things contain some special physical but equally mysterious stuff—*élan vital*) have been relegated to the trash heap of history, along with alchemy and astrology. Unless you are also prepared to declare that the world is flat and the sun is a fiery chariot pulled by winged horses—unless, in other words, your defiance of modern science is quite complete—you won't find any place to stand and fight for these obsolete ideas. So let's see what story can be told with the conservative resources of science. Maybe the idea that our minds evolved from simpler minds is not so bad after all.

Our macromolecule ancestors (and that's exactly and unmetaphorically what they were: our ancestors) were agent*like* in some ways, as the quotation from Eigen makes clear, and yet in other ways they were undeniably passive, floating randomly around, pushed hither and yon—waiting for action with their guns cocked, you might say, but not waiting *hopefully* or *resolutely* or *intently*. Their jaws might have gaped, but they were as mindless as a steel trap.

What changed? Nothing sudden. Before our ancestors got minds, they got bodies. First, they became simple cells, or prokaryotes, and eventually the prokaryotes took in some invaders, or boarders, and thereby became complex cells—the eukaryotes. By this time, roughly a billion years after the first appearance of simple cells, our ancestors were already

extraordinarily complex machines (made of machines made of machines), but they still didn't have minds. They were as passive and undirected in their trajectories as ever, but now they were equipped with many specialized subsystems, for extracting energy and material from the environment and protecting and repairing themselves when necessary.

The elaborate organization of all these coordinated parts was not very much like a mind. Aristotle had a name for it—or for its descendants: he called it a *nutritive soul*. A nutritive soul is not a thing; it is not, for instance, one of the microscopic subsystems floating around in the cytoplasm of a cell. It is a *principle of organization*; it is form, not substance, as Aristotle said. All living things—not only plants and animals but also unicellular organisms—have bodies that require a self-regulative and self-protective organization that can be differentially activated by different conditions. These organizations are brilliantly designed, by natural selection, and they are composed, at bottom, of lots of tiny passive switches that can be turned ON or OFF by equally passive conditions that the organisms encounter in their wanderings.

You yourself, like all other animals, have a nutritive soul—a self-regulative, self-protective organization—quite distinct from, and more ancient than, your nervous system: it consists of your metabolic system, your immune system, and the other staggeringly complex systems of self-repair and health maintenance in your body. The lines of communication used by these early systems were not nerves but blood vessels. Long before there were telephones and radios, there was the postal service, reliably if rather slowly transporting physical packages of valuable information around the world. And long before there were nervous systems in organisms, bodies relied on a low-tech postal system of sorts—the circulation of fluids within the body, reliably if rather slowly transporting valuable packages of information

to where they were needed for control and self-maintenance. We see the descendants of this primordial postal system in both animals and plants. In animals, the bloodstream carries goods and waste, but it has also been, since the early days, an information highway. The motion of fluids within plants also provides a relatively rudimentary medium for getting signals from one part of the plant to another. But in animals, we can see a major design innovation: the evolution of simple nervous systems—ancestors of the autonomic nervous system—capable of swifter and more efficient information transmission but still devoted, in the main, to internal affairs. An autonomic nervous system is not a mind at all but rather a control system, more along the lines of the nutritive soul of a plant, that preserves the basic integrity of the living system.

We sharply distinguish these ancient systems from our minds, and yet, curiously, the closer we look at the details of their operation the more mindlike we find them to be! The little switches are like primitive sense organs, and the effects that are produced when these switches are turned ON and OFF are like intentional actions. How so? In being effects produced by *information*-modulated, *goal*-seeking systems. It is *as if* these cells and cell assemblies were tiny, simple-minded *agents*, specialized servants rationally furthering their particular obsessive causes by acting in the ways their perception of circumstances dictated. The world is teeming with such entities, ranging from the molecular to the continental in size and including not only "natural" objects, such as plants, animals, and their parts (and the parts of their parts), but also many human artifacts. Thermostats, for instance, are a familiar example of such simple pseudo-agents.

I call all these entities, from the simplest to the most complex, *intentional systems*, and I call the perspective from which their agenthood (pseudo or genuine) is made visible, the *intentional stance*.

ADOPTING THE INTENTIONAL STANCE

The intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, whatever) by treating it *as if* it were a rational agent who governed its "choice" of "action" by a "consideration" of its "beliefs" and "desires." These terms in scare-quotes have been stretched out of their home use in what's often called "folk psychology," the everyday psychological discourse we use to discuss the mental lives of our fellow human beings. The intentional stance is the attitude or perspective we routinely adopt toward one another, so adopting the intentional stance toward something else seems to be deliberately *anthropomorphizing* it. How could this possibly be a good idea?

I will try to show that if done with care, adopting the intentional stance is not just a good idea but the key to unraveling the mysteries of the mind—all kinds of minds. It is a method that exploits similarities in order to discover differences—the huge collection of differences that have accumulated between the minds of our ancestors and ours, and also between our minds and those of our fellow inhabitants of the planet. It must be used with caution; we must walk a tightrope between vacuous metaphor on the one hand and literal falsehood on the other. Improper use of the intentional stance can seriously mislead the unwary researcher, but properly understood, it can provide a sound and fruitful perspective in several different fields, exhibiting underlying unity in the phenomena and directing our attention to the crucial experiments that need to be conducted.

The basic strategy of the intentional stance is to treat the entity in question as an agent, in order to predict—and thereby explain, in one sense—its actions or moves. The distinctive features of the intentional stance can best be seen by contrasting it with two more basic stances or strategies of

prediction: the *physical stance* and the *design stance*. The physical stance is simply the standard laborious method of the physical sciences, in which we use whatever we know about the laws of physics and the physical constitution of the things in question to devise our prediction. When I predict that a stone released from my hand will fall to the ground, I am using the physical stance. I don't attribute beliefs and desires to the stone; I attribute mass, or weight, to the stone, and rely on the law of gravity to yield my prediction. For things that are neither alive nor artifacts, the physical stance is the only available strategy, though it can be conducted at various levels of detail, from the subatomic to the astronomical. Explanations of why water bubbles when it boils, how mountain ranges come into existence, and where the energy in the sun comes from are explanations from the physical stance. Every physical thing, whether designed or alive or not, is subject to the laws of physics and hence behaves in ways that can be explained and predicted from the physical stance. If the thing I release from my hand is an alarm clock or a goldfish, I make the same prediction about its downward trajectory, on the same basis. And even a model airplane, or a bird, which may well take a different trajectory when released, behaves in ways that obey the laws of physics at every scale and at every moment.

Alarm clocks, being designed objects (unlike the rock), are also amenable to a fancier style of prediction—prediction from the design stance. The design stance is a wonderful shortcut, which we all use all the time. Suppose someone gives me a new digital alarm clock. It is a make and model quite novel to me, but a brief examination of its exterior buttons and displays convinces me that *if* I depress a few buttons just so, *then* some hours later the alarm clock will make a loud noise. I don't know what kind of noise it will be, but

it will be sufficient to awaken me. I don't need to work out the specific physical laws that explain this marvelous regularity; I don't need to take the thing apart, weighing its parts and measuring the voltages. I simply *assume* that it has a particular design—the design we call an alarm clock—and that it will function properly, as designed. I'm prepared to risk quite a lot on this prediction—not my life, perhaps, but my waking up in time to get to my scheduled lecture or catch a train. Design-stance predictions are riskier than physical-stance predictions, because of the extra assumptions I have to take on board: that an entity *is* designed as I suppose it to be, and that it will operate according to that design—that is, it will not malfunction. Designed things are occasionally misdesigned, and sometimes they break. But this moderate price I pay in riskiness is more than compensated by the tremendous ease of prediction. Design-stance prediction, when applicable, is a low-cost, low-risk shortcut, enabling me to finesse the tedious application of my limited knowledge of physics. In fact we all routinely risk our lives on design-stance predictions: we unhesitatingly plug in and turn on electrical appliances that could kill us if miswired; we voluntarily step into buses we know will soon accelerate us to lethal speeds; we calmly press buttons in elevators we have never been in before.

Design-stance prediction works wonderfully on well-designed artifacts, but it also works wonderfully on Mother Nature's artifacts—living things and their parts. Long before the physics and chemistry of plant growth and reproduction were understood, our ancestors quite literally bet their lives on the reliability of their design-stance knowledge of what seeds were *supposed* to do when planted. *If* I press a few seeds into the ground just so, *then* in a few months, with a modicum of further care from me, there will be food here to eat.

We have just seen that design-stance predictions are risky, compared with physical-stance predictions (which are safe but tedious to work out), and an even riskier and swifter stance is the intentional stance. It can be viewed, if you like, as a subspecies of the design stance, in which the designed thing is an agent of sorts. Suppose we apply it to the alarm clock. This alarm clock is my servant; if I *command it* to wake me up, by *giving it to understand* a particular time of awakening, I can rely on its internal ability to *perceive* when that time has arrived and dutifully execute the action it has promised. As soon as it comes to *believe* that the time for noise is NOW, it will be "motivated," thanks to my earlier instructions, to act accordingly. No doubt the alarm clock is so simple that this fanciful anthropomorphism is, strictly speaking, unnecessary for our understanding of why it does what it does—but notice that this is how we might explain to a child how to use an alarm clock: "You tell it when you want it to wake you up, and it remembers to do so, by making a loud noise."

Adoption of the intentional stance is more useful—indeed, well-nigh obligatory—when the artifact in question is much more complicated than an alarm clock. My favorite example is a chess-playing computer. There are hundreds of different computer programs that can turn a computer, whether it's a laptop or a supercomputer, into a chess player. For all their differences at the physical level and the design level, these computers all succumb neatly to the same simple strategy of interpretation: just think of them as rational agents who *want* to win, and who *know* the rules and principles of chess and the positions of the pieces on the board. Instantly your problem of predicting and interpreting their behavior is made vastly easier than it would be if you tried to use the physical or the design stance. At any moment in the chess game, simply look at the chessboard and draw up a list of all the legal moves available to the computer when it

is its turn to play (there will usually be several dozen candidates). Why restrict yourself to legal moves? Because, you reason, it wants to play winning chess and knows that it must make only legal moves to win, so, being rational, it restricts itself to these. Now rank the legal moves from best (wisest, most rational) to worst (stupidest, most self-defeating) and make your prediction: the computer will make the best move. You may well not be sure what the best move *is* (the computer may "appreciate" the situation better than you do!), but you can almost always eliminate all but four or five candidate moves, which still gives you tremendous predictive leverage.

Sometimes, when the computer finds itself in a tough predicament, with only one nonsuicidal move to make (a "forced" move), you can predict its move with supreme confidence. Nothing about the laws of physics forces this move, and nothing about the specific design of the computer forces this move. The move is forced by the overwhelmingly good *reasons* for making it and not any other move. Any chess player, constructed of whatever physical materials, would make it. Even a ghost or an angel would make it! You come up with your intentional-stance prediction on the basis of your bold assumption that *no matter how* the computer program has been designed, it has been designed well enough to be moved by such a good reason. You predict its behavior *as if* it were a rational agent.

The intentional stance is undeniably a useful shortcut in such a case, but how seriously should we take it? What does a computer care, really, about whether it wins or loses? Why say that the alarm clock *desires* to obey its master? We can use this contrast between natural and artificial goals to heighten our appreciation of the fact that all real goals ultimately spring from the predicament of a living, self-protective thing. But we must also recognize that the intentional stance *works* (when it does) whether or not the attributed

goals are genuine or natural or "really appreciated" by the so-called agent, and this tolerance is crucial to understanding how genuine goal-seeking could be established in the first place. Does the macromolecule *really* want to replicate itself? The intentional stance explains what is going on, regardless of how we answer that question. Consider a simple organism—say, a planarian or an amoeba—moving nonrandomly across the bottom of a laboratory dish, always heading to the nutrient-rich end of the dish, or away from the toxic end. This organism is seeking the good, or shunning the bad—*its own* good and bad, not those of some human artifact-user. Seeking one's own good is a fundamental feature of any rational agent, but are these simple organisms seeking or just "seeking?" We don't need to answer that question. The organism is a predictable intentional system in either case.

This is another way of making Socrates' point in the *Meno*, when he asks whether anyone ever knowingly desires evil. We intentional systems do sometimes desire evil, through misunderstanding or misinformation or sheer lunacy, but it is part and parcel of rationality to desire what is deemed good. It is this constitutive relationship between the good and the seeking of the good that is endorsed—or rather enforced—by the natural selection of our forebears: those with the misfortune to be genetically designed so that they seek what is bad for them leave no descendants in the long run. It is no accident that the products of natural selection seek (or "seek") what they deem (or "deem") to be good.

Even the simplest organisms, if they are to favor what is good for them, need some sense organs or discriminative powers—some simple switches that turn ON in the presence of good and OFF in its absence—and these switches, or *transducers*, must be united to the right bodily responses. This requirement is the birth of *function*. A rock can't malfunction, for it has not been well- or ill-equipped to further

any good. When we decide to interpret an entity from the intentional stance, it is as if we put ourselves in the role of its guardian, asking ourselves, in effect, "If *I* were in this organism's predicament, what would I do?" And here we expose the underlying anthropomorphism of the intentional stance: we treat all intentional systems as if they were just like us—which of course they are not.

Is this then a misapplication of our own perspective, the perspective *we mind-havers* share? Not necessarily. From the vantage point of evolutionary history, this is what has happened: Over billions of years, organisms gradually evolved, accumulating ever more versatile machinery designed to further their ever more complex and articulated goods. Eventually, with the evolution in our species of language and the varieties of reflectiveness that language permits (a topic for later chapters), we emerged with the ability to wonder the wonders with which we began this book—wonders about the minds of other entities. These wonders, naively conducted by our ancestors, led to *animism*, the idea that each moving thing has a mind or soul (*anima,* in Latin). We began to ask ourselves not only whether the tiger wanted to eat us—which it probably did—but why the rivers wanted to reach the seas, and what the clouds wanted from us in return for the rain we asked of them. As we became more sophisticated—and this is a very recent historical development, not anything to be discerned in the vast reaches of evolutionary time—we gradually withdrew the intentional stance from what we now call *inanimate* nature, reserving it for things more like us: animals, in the main, but also plants under many conditions. We still "trick" flowers into blooming prematurely by "deceiving" them with artificial spring warmth and light, and "encourage" vegetables to send down longer roots by withholding from them the water they want so badly. (A logger once explained to me how he knew we would find no white pines among the trees in some high

ground in my forest—"Pines like to keep their feet wet.") This way of thinking about plants is not only natural and harmless but positively an aid to comprehension and an important lever for discovery. When biologists discover that a plant has some rudimentary discriminatory organ, they immediately ask themselves what the organ is for—what devious project does the plant have that requires it to obtain information from its environment on this topic? Very often the answer is an important scientific discovery.

Intentional systems are, by definition, all and only those entities whose behavior is predictable/explicable from the intentional stance. Self-replicating macromolecules, thermostats, amoebas, plants, rats, bats, people, and chess-playing computers are all intentional systems—some much more interesting than others. Since the point of the intentional stance is to treat an entity as an agent in order to predict its actions, we have to suppose that it is a smart agent, since a stupid agent might do any dumb thing at all. This bold leap of supposing that the agent will make only the smart moves (given its limited perspective) is what gives us the leverage to make predictions. We describe that limited perspective by attributing *particular* beliefs and desires to the agent on the basis of its perception of the situation and its goals or needs. Since our predictive leverage in this exercise is critically dependent on this particularity—since it is sensitive to the particular way the beliefs and desires are expressed by us, the theorists, or represented by the intentional system in question, I call such systems *intentional* systems. They exhibit what philosophers call *intentionality*.

"Intentionality," in this special philosophical sense, is such a controversial concept, and is so routinely misunderstood and misused by nonphilosophers, that I must pause to belabor its definition. Unfortunately for interdisciplinary communication, the philosophical term "intentionality" has two false friends—perfectly good words that are readily con-

fused with it, and indeed are rather closely related to it. One is an ordinary term, the other is technical (and I will postpone its introduction briefly). In ordinary parlance, we often discuss whether someone's action was intentional or not. When the driver crashed into the bridge abutment, was he intentionally committing suicide, or had he fallen asleep? When you called the policeman "Dad" just then, was that intentional, or a slip of the tongue? Here we are asking, are we not, about the intentionality of the two deeds? Yes, in the ordinary sense; no, in the philosophical sense.

Intentionality in the philosophical sense is just *aboutness*. Something exhibits intentionality if its competence is in some way *about* something else. An alternative would be to say that something that exhibits intentionality contains a *representation* of something else—but I find that less revealing and more problematic. Does a lock contain a representation of the key that opens it? A lock and key exhibit the crudest form of intentionality; so do the opioid receptors in brain cells—receptors that are designed to accept the endorphin molecules that nature has been providing in brains for millions of years. Both can be tricked—that is, opened by an impostor. Morphine molecules are artifactual skeleton keys that have recently been fashioned to open the opioid-receptor doors too. (In fact it was the discovery of these highly specific receptors which inspired the search that led to the discovery of endorphins, the brain's own painkillers. There must have been something already present in the brain, reasearchers reasoned, for these specialized receptors to have been *about* in the first place.) This lock-and-key variety of crude aboutness is the basic design element out of which nature has fashioned the fancier sorts of subsystems that may more deservedly be called representation systems, so we will have to analyze the aboutness of these representations in terms of the (quasi?) aboutness of locks-and-keys in any case. We can stretch a point and say

that the present shape of the bimetallic spring in a thermostat is a representation of the present room temperature, and that the position of the thermostat's adjustable lever is a representation of the desired room temperature, but we can equally well deny that these are, properly speaking, representations. They do, however, embody information *about* room temperature, and it is by virtue of that embodiment that they contribute to the competence of a simple intentional system.

Why do philosophers call aboutness "intentionality"? It all goes back to the medieval philosophers who coined the term, noting the similarity between such phenomena and the act of aiming an arrow at something (*intendere arcum in*). Intentional phenomena are equipped with metaphorical arrows, you might say, aimed at something or other—at whatever it is the phenomena are about or refer to or allude to. But of course many phenomena that exhibit this minimal sort of intentionality do not do anything *intentionally,* in the everyday sense of the term. Perceptual states, emotional states, and states of memory, for example, all exhibit aboutness without necessarily being intentional in the ordinary sense; they can be entirely involuntary or automatic responses to one thing or another. There is nothing intentional about recognizing a horse when it looms into view, but your state of recognition exhibits very particular aboutness: you recognize it *as* a horse. If you had misperceived it *as* a moose or a man on a motorcycle, your perceptual state would have had a different aboutness. It would have aimed its arrow rather differently—at something nonexistent, in fact, but nevertheless quite definite: either the moose that never was, or the illusory motorcyclist. There is a large psychological difference between mistakenly thinking you're in the presence of a moose and mistakenly thinking you're in the presence of a man on a motorcycle, a difference with predictable consequences. The medieval theorists noted that

the arrow of intentionality could thus be aimed at nothing while nevertheless being aimed in a rather particular way. They called the object of your thought, real or not, the *intentional object*.

In order to think about something, you must have a way—one way among many possible ways—of thinking about it. Any intentional system is dependent on its particular ways of thinking about—perceiving, searching for, identifying, fearing, recalling—whatever it is that its "thoughts" are about. It is this dependency that creates all the opportunities for confusion, both practical and theoretical. Practically, the best way to confuse a particular intentional system is to exploit a flaw in its way(s) of perceiving or thinking about whatever it needs to think about. Nature has explored countless variations on this theme, since confusing other intentional systems is a major goal in the life of most intentional systems. After all, one of the primary desires of any living intentional system is the desire for the food needed to fuel growth, self-repair, and reproduction, so every living thing needs to distinguish the food (the good material) from the rest of the world. It follows that another primary desire is to avoid becoming the food of another intentional system. So camouflage, mimicry, stealth, and a host of other stratagems have put nature's locksmiths to the test, provoking the evolution of ever more effective ways of distinguishing one thing from another and keeping track of them. But no way is ever foolproof. There is no *taking* without the possibility of *mistaking*. That's why it's so important for us as theorists to be able to identify and distinguish the different varieties of taking (and mistaking) that can occur in intentional systems. In order to make sense of a system's actual "take" on its circumstances, we have to have an accurate picture of its dependence on its particular capacities for distinguishing things—its ways of "thinking about" things.

Unfortunately, however, as theorists we have tended to overdo it, treating *our own* well-nigh limitless capacity for distinguishing one thing from another in our thoughts (thanks to our ability to use language) as if it were the hallmark of all genuine intentionality, all aboutness worthy of the name. For instance, when a frog's tongue darts out and catches whatever is flying by, the frog may make a mistake—it may ingest a ball bearing thrown by a mischievous child, or a fisherman's lure on a monofilament thread, or some other inedible anomaly. The frog has made a mistake, but *exactly* which mistake(s) has it made? What did the frog "think" it was grabbing? A fly? Airborne food? A moving dark convexity? We language users can draw indefinitely fine distinctions of content for the candidate frog-thought, and there has been an unexamined assumption that before we can attribute any *real* intentionality to the frog we have to narrow down the content of the frog's states and acts with the same precision we can use (in principle) when we consider human thoughts and their propositional content.

This has been a major source of theoretical confusion, and to make matters worse, there is a handy technical term from logic that refers to just this capacity of language for making indefinitely fine-grained discriminations: *intensionality*. With an *s*. Intensionality-with-an-s is a feature of languages; it has no direct application to any other sort of representational system (pictures, maps, graphs, "search images," . . . *minds*). According to standard usage among logicians, the words or symbols in a language can be divided into the logical, or function, words ("if," "and," "or," "not," "all," "some," . . .) and the *terms* or *predicates*, which can be as various as the topic of discussion ("red," "tall," "grandfather," "oxygen," "second-rate composer of sonnets," . . .). Every meaningful term or predicate of a language has an *extension*—the thing or set of things to which the term refers—and an *intension*—the particular way in which this

thing or set of things is picked out or determined. "Chelsea Clinton's father" and "president of the United States in 1995" name the very same thing—Bill Clinton—and hence have the same extension, but they zero in on this common entity in different ways, and hence have difference intensions. The term "equilateral triangle" picks out exactly the same set of things as the term "equiangular triangle," so these two terms have the same extension, but clearly they don't mean the same thing: one term is *about* a triangle's sides being equal and the other is *about* the angles being equal. So intension (with an s) is contrasted to extension, and means, well, *meaning*. And isn't that what intentionality-with-a-t means, too?

For many purposes, logicians note, we can ignore differences in the *intensions* of terms and just keep track of *extensions*. After all, a rose by any other name would smell as sweet, so if roses are the topic, the indefinitely many different ways of getting the class of roses into the discussion should be equivalent, from a logical point of view. Since water *is* H_2O, anything truly said of water, using the term "water," will be just as truly said if we substitute the term "H_2O" in its place—even if these two terms are subtly different in meaning, or intension. This freedom is particularly obvious and useful in such topic areas as mathematics, where you can always avail yourself of the practice of "substituting equals for equals," replacing "4^2" by "16" or vice versa, since these two different terms refer to one and the same number. Such freedom of substitution within linguistic contexts is aptly called *referential transparency*: you can see right through the terms, in effect, to the things the terms refer to. But when the topic is not roses but *thinking-about-roses*, or *talking-about-(thinking-about)-roses*, differences in intension can matter. So whenever the topic is intentional systems and their beliefs and desires, the language used by the theorist is intension-sensitive. A logician would say that

such discourse exhibits *referential opacity*; it is not transparent; the terms themselves get in the way and interfere in subtle and confusing ways with the topic.

To see how referential opacity actually matters when we adopt the intentional stance, let's consider a root case of the intentional stance in action, applied to a human being. We do this effortlessly every day, and seldom spell out what is involved, but here is an example, drawn from a recent philosophical article—an example that goes rather weirdly but usefully into more detail than usual:

> Brutus wanted to kill Caesar. He believed that Caesar was an ordinary mortal, and that, given this, stabbing him (by which we mean plunging a knife into his heart) was a way of killing him. He thought that he could stab Caesar, for he remembered that he had a knife and saw that Caesar was standing next to him on his left in the Forum. So Brutus was motivated to stab the man to his left. He did so, thereby killing Caesar. (Israel, Perry, and Tutiya, 1993. p. 515)

Notice that the term "Caesar" is surreptitiously playing a crucial double role in this explanation—not just in the normal, transparent way of picking out a man, Caesar, the chap in the toga standing in the Forum, but in picking out the man *in the way Brutus himself picks him out.* It is not enough for Brutus to see Caesar standing next to him; he has to see that he *is* Caesar, the man he wants to kill. If Brutus mistook Caesar, the man to his left, for Cassius, then he wouldn't try to kill him: he wouldn't have been motivated, as the authors say, to stab the man to his left, since he would not have drawn the crucial connection in his mind—the link identifying the man to his left with his goal.