CHAPTER 1

PROLOGUE

Everything should be made as simple as possible,
but not simpler.

—ALBERT EINSTEIN

This book tries to explain how minds work. How can intelligence emerge from nonintelligence? To answer that, we'll show that you can build a mind from many little parts, each mindless by itself.

I'll call "Society of Mind" this scheme in which each mind is made of many smaller processes. These we'll call *agents*. Each mental agent by itself can only do some simple thing that needs no mind or thought at all. Yet when we join these agents in societies—in certain very special ways—this leads to true intelligence.

There's nothing very technical in this book. It, too, is a society—of many small ideas. Each by itself is only common sense, yet when we join enough of *them* we can explain the strangest mysteries of mind.

One trouble is that these ideas have lots of cross-connections. My explanations rarely go in neat, straight lines from start to end. I wish I could have lined them up so that you could climb straight to the top, by mental stair-steps, one by one. Instead they're tied in tangled webs.

Perhaps the fault is actually mine, for failing to find a tidy base of neatly ordered principles. But I'm inclined to lay the blame upon the nature of the mind: much of its power seems to stem from just the messy ways its agents cross-connect. If so, that complication can't be helped; it's only what we must expect from evolution's countless tricks.

What can we do when things are hard to describe? We start by sketching out the roughest shapes to serve as scaffolds for the rest; it doesn't matter very much if some of those forms turn out partially wrong. Next, draw details to give these skeletons more lifelike flesh. Last, in the final filling-in, discard whichever first ideas no longer fit.

That's what we do in real life, with puzzles that seem very hard. It's much the same for shattered pots as for the cogs of great machines. Until you've seen some of the rest, you can't make sense of any part.

1.1 THE AGENTS OF THE MIND

Good theories of the mind must span at least three different scales of time: slow, for the billion years in which our brains have evolved; fast, for the fleeting weeks and months of infancy and childhood; and in between, the centuries of growth of our ideas through history.

To explain the mind, we have to show how minds are built from mindless stuff, from parts that are much smaller and simpler than anything we'd consider smart. Unless we can explain the mind in terms of things that have no thoughts or feelings of their own, we'll only have gone around in a circle. But what could those simpler particles be—the "agents" that compose our minds? This is the subject of our book, and knowing this, let's see our task. There are many questions to answer.

Function:	*How do agents work?*
Embodiment:	*What are they made of?*
Interaction:	*How do they communicate?*
Origins:	*Where do the first agents come from?*
Heredity:	*Are we all born with the same agents?*
Learning:	*How do we make new agents and change old ones?*
Character:	*What are the most important kinds of agents?*
Authority:	*What happens when agents disagree?*
Intention:	*How could such networks want or wish?*
Competence:	*How can groups of agents do what separate agents cannot do?*
Selfness:	*What gives them unity or personality?*
Meaning:	*How could they understand anything?*
Sensibility:	*How could they have feelings and emotions?*
Awareness:	*How could they be conscious or self-aware?*

How could a theory of the mind explain so many things, when every separate question seems too hard to answer by itself? These questions all seem difficult, indeed, when we sever each one's connections to the other ones. But once we see the mind as a society of agents, each answer will illuminate the rest.

1.2 THE MIND AND THE BRAIN

> *It was never supposed* [the poet Imlac said] *that cogitation is inherent in matter, or that every particle is a thinking being. Yet if any part of matter be devoid of thought, what part can we suppose to think? Matter can differ from matter only in form, bulk, density, motion and direction of motion: to which of these, however varied or combined, can consciousness be annexed? To be round or square, to be solid or fluid, to be great or little, to be moved slowly or swiftly one way or another, are modes of material existence, all equally alien from the nature of cogitation. If matter be once without thought, it can only be made to think by some new modification, but all the modifications which it can admit are equally unconnected with cogitative powers.*
>
> —SAMUEL JOHNSON

How could solid-seeming brains support such ghostly things as thoughts? This question troubled many thinkers of the past. The world of thoughts and the world of things appeared to be too far apart to interact in any way. So long as thoughts seemed so utterly different from everything else, there seemed to be no place to start.

A few centuries ago it seemed equally impossible to explain Life, because living things appeared to be so different from anything else. Plants seemed to grow from nothing. Animals could move and learn. Both could reproduce themselves—while nothing else could do such things. But then that awesome gap began to close. Every living thing was found to be composed of smaller cells, and cells turned out to be composed of complex but comprehensible chemicals. Soon it was found that plants did not create any substance at all but simply extracted most of their material from gases in the air. Mysteriously pulsing hearts turned out to be no more than mechanical pumps, composed of networks of muscle cells. But it was not until the present century that John von Neumann showed theoretically how cell-machines could reproduce while, almost independently, James Watson and Francis Crick discovered how each cell actually makes copies of its own hereditary code. No longer does an educated person have to seek any special, vital force to animate each living thing.

Similarly, a century ago, we had essentially no way to start to explain how thinking works. Then psychologists like Sigmund Freud and Jean Piaget produced their theories about child development. Somewhat later, on the mechanical side, mathematicians like Kurt Gödel and Alan Turing began to reveal the hitherto unknown range of what machines could be made to do. These two streams of thought began to merge only in the 1940s, when Warren McCulloch and Walter Pitts began to show how machines might be made to see, reason, and remember. Research in the modern science of Artificial Intelligence started only in the 1950s, stimulated by the invention of modern computers. This inspired a flood of new ideas about how machines could do what only minds had done previously.

Most people still believe that no machine could ever be conscious, or feel ambition, jealousy, humor, or have any other mental life-experience. To be sure, we are still far from being able to create machines that do all the things people do. But this only means that we need better theories about how thinking works. This book will show how the tiny machines that we'll call "agents of the mind" could be the long sought "particles" that those theories need.

1.3 THE SOCIETY OF MIND

You know that everything you think and do is thought and done by you. But what's a "you"? What kinds of smaller entities cooperate inside your mind to do your work? To start to see how minds are like societies, try this: *pick up a cup of tea!*

> *Your GRASPING agents want to keep hold of the cup.*
> *Your BALANCING agents want to keep the tea from spilling out.*
> *Your THIRST agents want you to drink the tea.*
> *Your MOVING agents want to get the cup to your lips.*

Yet none of these consume your mind as you roam about the room talking to your friends. You scarcely think at all about *Balance*; *Balance* has no concern with *Grasp*; *Grasp* has no interest in *Thirst*; and *Thirst* is not involved with your social problems. Why not? Because they can depend on one another. If each does its own little job, the really big job will get done by all of them together: drinking tea.

How many processes are going on, to keep that teacup level in your grasp? There must be at least a hundred of them, just to shape your wrist and palm and hand. Another thousand muscle systems must work to manage all the moving bones and joints that make your body walk around. And to keep everything in balance, each of those processes has to communicate with some of the others. What if you stumble and start to fall? Then many other processes quickly try to get things straight. Some of them are concerned with how you lean and where you place your feet. Others are occupied with what to do about the tea: you wouldn't want to burn your own hand, but neither would you want to scald someone else. You need ways to make quick decisions.

All this happens while you talk, and none of it appears to need much thought. But when you come to think of it, neither does your talk itself. What kinds of agents choose your words so that you can express the things you mean? How do those words get arranged into phrases and sentences, each connected to the next? What agencies inside your mind keep track of all the things you've said—and, also, whom you've said them to? How foolish it can make you feel when you repeat—unless you're sure your audience is new.

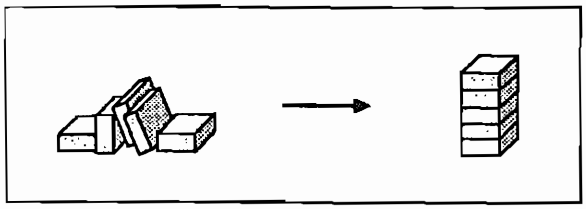
We're always doing several things at once, like planning and walking and talking, and this all seems so natural that we take it for granted. But these processes actually involve more machinery than anyone can understand all at once. So, in the next few sections of this book, we'll focus on just one ordinary activity—*making things with children's building-blocks*. First we'll break this process into smaller parts, and then we'll see how each of them relates to all the other parts.

In doing this, we'll try to imitate how Galileo and Newton learned so much by studying the simplest kinds of pendulums and weights, mirrors and prisms. Our study of how to build with blocks will be like focusing a microscope on the simplest objects we can find, to open up a great and unexpected universe. It is the same reason why so many biologists today devote more attention to tiny germs and viruses than to magnificent lions and tigers. For me and a whole generation of students, the world of work with children's blocks has been the prism and the pendulum for studying intelligence.

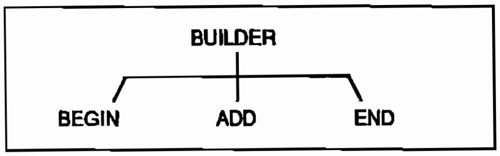
In science, one can learn the most by studying what seems the least.

1.4 THE WORLD OF BLOCKS

Imagine a child playing with blocks, and imagine that this child's mind contains a host of smaller minds. Call them mental agents. Right now, an agent called *Builder* is in control. *Builder*'s specialty is making towers from blocks.

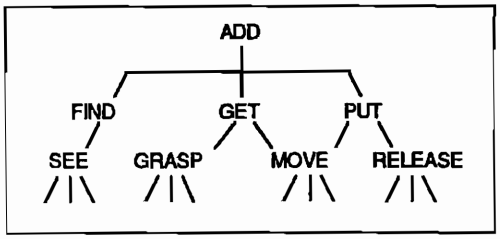


Our child likes to watch a tower grow as each new block is placed on top. But building a tower is too complicated a job for any single, simple agent, so *Builder* has to ask for help from several other agents:



Choose a place to start the tower.
Add a new block to the tower.
Decide whether it is high enough.

In fact, even to find another block and place it on the tower top is too big for a job for any single agent. So *Add*, in turn, must call for other agents' help. Before we're done, we'll need more agents than would fit in any diagram.



First ADD must FIND a new block.
Then the hand must GET that
block and PUT it on the tower top.

Why break things into such small parts? Because minds, like towers, are made that way—except that they're composed of processes instead of blocks. And if making stacks of blocks seems insignificant—remember that you didn't always feel that way. When first you found some building toys in early childhood, you probably spent joyful weeks of learning what to do with them. If such toys now seem relatively dull, then you must ask yourself how *you* have changed. Before you turned to more ambitious things, it once seemed strange and wonderful to be able to build a tower or a house of blocks. Yet, though all grown-up persons know how to do such things, *no one understands how we learn to do them!* And *that* is what will concern us here. To pile up blocks into heaps and rows: these are skills each of us learned so long ago that we can't remember learning them at all. Now they seem mere common sense—and that's what makes psychology hard. This forgetfulness, the amnesia of infancy, makes us assume that all our wonderful abilities were always there inside our minds, and we never stop to ask ourselves how they began and grew.

1.5 COMMON SENSE

You cannot think about thinking, without thinking about thinking about something.
—Seymour Papert

We found a way to make our tower builder out of parts. But *Builder* is really far from done. To build a simple stack of blocks, our child's agents must accomplish all these other things.

> ***See** must recognize its blocks, whatever their color, size, and place—in spite of different backgrounds, shades, and lights, and even when they're partially obscured by other things.*
>
> *Then, once that's done, **Move** has to guide the arm and hand through complicated paths in space, yet never strike the tower's top or hit the child's face.*
>
> *And think how foolish it would seem, if **Find** were to see, and **Grasp** were to grasp, a block supporting the tower top!*

When we look closely at these requirements, we find a bewildering world of complicated questions. For example, how could *Find* determine which blocks are still available for use? *It would have to "understand" the scene in terms of what it is trying to do.* This means that we'll need theories both about what it means to understand and about how a machine could have a goal. Consider all the *practical* judgments that an actual *Builder* would have to make. It would have to decide whether there are enough blocks to accomplish its goal and whether they are strong and wide enough to support the others that will be placed on them.

What if the tower starts to sway? A real builder must guess the cause. It is because some joint inside the column isn't square enough? Is the foundation insecure, or is the tower too tall for its width? Perhaps it is only because the last block was placed too roughly.

All children learn about such things, but we rarely ever think about them in our later years. By the time we are adults we regard all of this to be simple "common sense." But that deceptive pair of words conceals almost countless different skills.

> *Common sense is not a simple thing. Instead, it is an immense society of hard-earned practical ideas—of multitudes of life-learned rules and exceptions, dispositions and tendencies, balances and checks.*

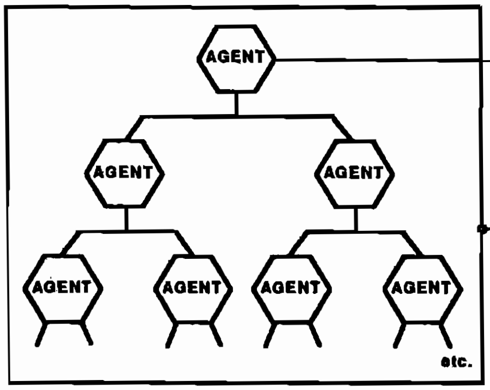
If common sense is so diverse and intricate, what makes it seem so obvious and natural? This illusion of simplicity comes from losing touch with what happened during infancy, when we formed our first abilities. As each new group of skills matures, we build more layers on top of them. As time goes on, the layers below become increasingly remote until, when we try to speak of them in later life, we find ourselves with little more to say than "*I don't know*."

1.6 AGENTS AND AGENCIES

We want to explain intelligence as a combination of simpler things. This means that we must be sure to check, at every step, that none of our agents is, itself, intelligent. Otherwise, our theory would end up resembling the nineteenth-century "chessplaying machine" that was exposed by Edgar Allan Poe to actually conceal a human dwarf inside. Accordingly, whenever we find that an agent has to do anything complicated, we'll replace it with a subsociety of agents that do simpler things. Because of this, the reader must be prepared to feel a certain sense of loss. When we break things down to their smallest parts, they'll each seem dry as dust at first, as though some essence has been lost.

For example, we've seen how to construct a tower-building skill by making *Builder* from little parts like *Find* and *Get.* Now, where does its "knowing-how-to-build" reside when, clearly, it is not in any part—and yet those parts are all that *Builder* is? The answer: It is not enough to explain only what each separate agent does. We must also understand how those parts are interrelated—that is, how *groups* of agents can accomplish things.

Accordingly, each step in this book uses two different ways to think about agents. If you were to watch *Builder* work, from the outside, with no idea of how it works inside, you'd have the impression that it knows how to build towers. But if you could see *Builder* from the inside, you'd surely find no knowledge there. You would see nothing more than a few switches, arranged in various ways to turn each other on and off. *Does **Builder** "really know" how to build towers?* The answer depends on how you look at it. Let's use two different words, "*agent*" and "*agency*," to say why *Builder* seems to lead a double life. As agency, it seems to know its job. As agent, it cannot know anything at all.



Seen by itself, as an agent, BUILDER is just a simple process that turns other agents on and off.

Seen from outside, as an agency, BUILDER does whatever all its subagents accomplish, using one another's help.

When you drive a car, you regard the steering wheel as an agency that you can use to change the car's direction. You don't care how it works. But when something goes wrong with the steering, and you want to understand what's happening, it's better to regard the steering wheel as just one agent in a larger agency: it turns a shaft that turns a gear to pull a rod that shifts the axle of a wheel. Of course, one doesn't always want to take this microscopic view; if you kept all those details in mind while driving, you might crash because it took too long to figure out which way to turn the wheel. Knowing how is not the same as knowing why. In this book, we'll always be switching between agents and agencies because, depending on our purposes, we'll have to use different viewpoints and kinds of descriptions.

CHAPTER 2

WHOLES AND PARTS

It is the nature of the mind that makes individuals kin, and the differences in the shape, form, or manner of the material atoms out of whose intricate relationships that mind is built are altogether trivial.

—ISAAC ASIMOV

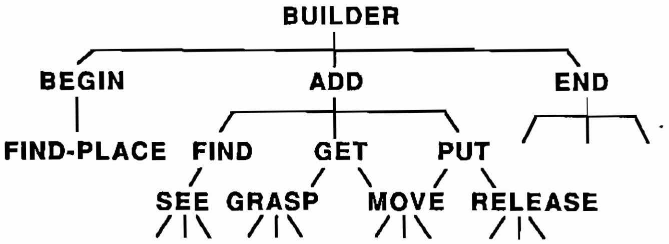
2.1 COMPONENTS AND CONNECTIONS

We saw that *Builder's* skill could be reduced to the simpler skills of *Get* and *Put*. Then we saw how these, in turn, could be made of even simpler ones. *Get* merely needs to *Move* the hand to *Grasp* the block that *Find* just found. *Put* only has to *Move* the hand so that it puts that block upon the tower top. So it might appear that all of *Builder's* functions have been "reduced" to things that simpler parts can do.

But something important has been left out. *Builder* is not merely a collection of parts like *Find*, *Get*, *Put*, and all the rest. For *Builder* would not work at all unless those agents were linked to one another by a suitable network of interconnections.

PUT	BEGIN
ADD	END
SEE	FIND
GET	MOVE
GRASP	RELEASE
FIND-PLACE	

Agents by Themselves



Agents in a Bureaucracy

Could you predict what *Builder* does from knowing just that left-hand list? Of course not: you must also know which agents work for which. Similarly, you couldn't predict what would happen in a human community from knowing only what each separate individual can do; you must also know how they are organized—that is, who talks to whom. And it's the same for understanding any large and complex thing. First, we must know how each separate part works. Second, we must know how each part interacts with those to which it is connected. And third, we have to understand how all these local interactions combine to accomplish what that system *does*—as seen from the outside.

In the case of the human brain, it will take a long time to solve these three kinds of problems. First we will have to understand how brain cells work. This will be difficult because there are hundreds of different types of brain cells. Then we'll have to understand how the cells of each type interact with the other types of cells to which they connect. There could be thousands of these different kinds of interactions. Then, finally, comes the hardest part: we'll also have to understand how our billions of brain cells are organized into societies. To do this, we'll need to develop many new theories and organizational concepts. The more we can find out about how our brains evolved from those of simpler animals, the easier that task will be.

2.2 NOVELISTS AND REDUCTIONISTS

It's always best when mysteries can be explained in terms of things we know. But when we find this hard to do, we must decide whether to keep trying to make old theories work or to discard them and try new ones. I think this is partly a matter of personality. Let's call "Reductionists" those people who prefer to build on old ideas, and "Novelists" the ones who like to champion new hypotheses. Reductionists are usually right—at least at science's cautious core, where novelties rarely survive for long. Outside that realm, though, novelists reign, since older ideas have had more time to show their flaws.

It really is amazing how certain sciences depend upon so few kinds of explanations. The science of physics can now explain virtually everything we see, *at least in principle*, in terms of how a very few kinds of particles and force-fields interact. Over the past few centuries reductionism has been remarkably successful. What makes it possible to describe so much of the world in terms of so few basic rules? No one knows.

Many scientists look on chemistry and physics as ideal models of what psychology should be like. After all, the atoms in the brain are subject to the same all-inclusive physical laws that govern every other form of matter. Then can we also explain what our brains actually do entirely in terms of those same basic principles? The answer is no, simply because even if we understood how each of our billions of brain cells work separately, this would not tell us how the brain works as an agency. The "laws of thought" depend not only upon the properties of those brain cells, but also on how they are connected. And these connections are established not by the basic, "general" laws of physics, but by the particular arrangements of the millions of bits of information in our inherited genes. To be sure, "general" laws apply to everything. But, for that very reason, they can rarely explain anything in particular.

Does this mean that psychology must reject the laws of physics and find its own? Of course not. It is not a matter of *different* laws, but of *additional* kinds of theories and principles that operate at higher levels of organization. Our ideas of how *Builder* works as an agency need not, and must not, conflict with our knowledge of how *Builder*'s lower-level agents work. Each higher level of description must *add* to our knowledge about lower levels, rather than replace it. We'll return to the idea of "level" at many places in this book.

Will psychology ever resemble any of the sciences that have successfully reduced their subjects to only a very few principles? That depends on what you mean by "few." In physics, we're used to explanations in terms of perhaps a dozen basic principles. For psychology, our explanations will have to combine hundreds of smaller theories. To physicists, that number may seem too large. To humanists, it may seem too small.

2.3 PARTS AND WHOLES

We're often told that certain wholes are "more than the sum of their parts." We hear this expressed with reverent words like "holistic" and "gestalt," whose academic tones suggest that they refer to clear and definite ideas. But I suspect the actual function of such terms is to anesthetize a sense of ignorance. We say "gestalt" when things combine to act in ways we can't explain, "holistic" when we're caught off guard by unexpected happenings and realize we understand less than we thought we did. For example, consider the two sets of questions below, the first "subjective" and the second "objective":

What makes a drawing more than just its separate lines?
How is a personality more than a set of traits?
In what way is a culture more than a mere collection of customs?

What makes a tower more than separate blocks?
Why is a chain more than its various links?
How is a wall more than a set of many bricks?

Why do the "objective" questions seem less mysterious? Because we have good ways to answer them—in terms of how things interact. To explain how walls and towers work, we just point out how every block is held in place by its neighbors and by gravity. To explain why chain-links cannot come apart, we can demonstrate how each would get in its neighbors' way. These explanations seem almost self-evident to adults. However, they did not seem so simple when we were children, and it took each of us several years to learn how real-world objects interact—for example, to prevent any two objects from ever being in the same place. We regard such knowledge as "obvious" only because we cannot remember how hard it was to learn.

Why does it seem so much harder to explain our reactions to drawings, personalities, and cultural traditions? Many people assume that those "subjective" kinds of questions are impossible to answer because they involve our minds. But that doesn't mean they can't be answered. It only means that we must first know more about our minds.

"Subjective" reactions are also based on how things interact. The difference is that here we are not concerned with objects in the world outside, but with processes inside our brains.

In other words, those questions about arts, traits, and styles of life are actually quite technical. They ask us to explain what happens among the agents in our minds. But this is a subject about which we have never learned very much—and neither have our sciences. Such questions will be answered in time. But it will just prolong the wait if we keep using pseudo-explanation words like "holistic" and "gestalt." True, sometimes giving names to things can help by leading us to focus on some mystery. It's harmful, though, when naming leads the mind to think that names alone bring meaning close.

2.4 HOLES AND PARTS

It has been the persuasion of an immense majority of human beings that sensibility and thought [as distinguished from matter] *are, in their own nature, less susceptible of division and decay, and that, when the body is resolved into its elements, the principle which animated it will remain perpetual and unchanged. However, it is probable that what we call thought is not an actual being, but no more than the relation between certain parts of that infinitely varied mass, of which the rest of the universe is composed, and which ceases to exist as soon as those parts change their position with respect to each other.*

—PERCY BYSSHE SHELLEY

What is Life? One dissects a body but finds no life inside. What is Mind? One dissects a brain but finds no mind therein. Are life and mind so much more than the "sum of their parts" that it is useless to search for them? To answer that, consider this parody of a conversation between a Holist and an ordinary Citizen.

> **Holist:** *"I'll prove no box can hold a mouse. A box is made by nailing six boards together. But it's obvious that no box can hold a mouse unless it has some 'mouse-tightness' or 'containment.'* Now, *no single board contains any containment, since the mouse can just walk away from it. And if there is no containment in one board, there can't be any in six boards. So the box can have no mousetightness at all. Theoretically, then, the mouse can escape!"*
>
> **Citizen:** *"Amazing. Then what* does *keep a mouse in a box?"*
>
> **Holist:** *"Oh, simple. Even though it has no real mousetightness, a good box can 'simulate' it so well that the mouse is fooled and can't figure out how to escape."*

What, then, keeps the mouse confined? Of course, it is the way a box prevents motion in all directions, because each board bars escape in a certain direction. The left side keeps the mouse from going left, the right from going right, the top keeps it from leaping out, and so on. The secret of a box is simply in how the boards are arranged to prevent motion in *all* directions! That's what *containing* means. So it's silly to expect any separate board by itself to contain any *containment*, even though each contributes to the containing. It is like the cards of a straight flush in poker: only the full hand has any value at all.

The same applies to words like *life* and *mind.* It is foolish to use these words for describing the smallest components of living things because these words were invented to describe how larger assemblies interact. Like *boxing-in*, words like *living* and *thinking* are useful for describing phenomena that result from certain combinations of relationships. The reason *box* seems nonmysterious is that everyone understands how the boards of a well-made box interact to prevent motion in any direction. In fact, the word *life* has already lost most of its mystery—at least for modern biologists, because they understand so many of the important interactions among the chemicals in cells. But *mind* still holds its mystery—because we still know so little about how mental agents interact to accomplish all the things they do.

2.5 EASY THINGS ARE HARD

In the late 1960s *Builder* was embodied in the form of a computer program at the MIT Artificial Intelligence Laboratory. Both my collaborator, Seymour Papert, and I had long desired to combine a mechanical hand, a television eye, and a computer into a robot that could build with children's building-blocks. It took several years for us and our students to develop *Move, See, Grasp,* and hundreds of other little programs we needed to make a working *Builder*-agency. I like to think that this project gave us glimpses of what happens inside certain parts of children's minds when they learn to "play" with simple toys. The project left us wondering if even a thousand microskills would be enough to enable a child to fill a pail with sand. It was this body of experience, more than anything we'd learned about psychology, that led us to many ideas about societies of mind.

To do those first experiments, we had to build a mechanical Hand, equipped with sensors for pressure and touch at its fingertips. Then we had to interface a television camera with our computer and write programs with which that Eye could discern the edges of the building-blocks. It also had to recognize the Hand itself. When those programs didn't work so well, we added more programs that used the fingers' feeling-sense to verify that things were where they visually seemed to be. Yet other programs were needed to enable the computer to move the Hand from place to place while using the Eye to see that there was nothing in its way. We also had to write higher-level programs that the robot could use for planning what to do—and still more programs to make sure that those plans were actually carried out. To make this all work reliably, we needed programs to verify at every step (again by using Eye and Hand) that what had been planned inside the mind did actually take place outside—or else to correct the mistakes that occurred.

In attempting to make our robot work, we found that many everyday problems were much more complicated than the sorts of problems, puzzles, and games adults consider hard. At every point, in that world of blocks, when we were forced to look more carefully than usual, we found an unexpected universe of complications. Consider just the seemingly simple problem of not reusing blocks already built into the tower. To a person, this seems simple common sense: *"Don't use an object to satisfy a new goal if that object is already involved in accomplishing a prior goal."* No one knows exactly how human minds do this. Clearly we learn from experience to recognize the situations in which difficulties are likely to occur, and when we're older we learn to plan ahead to avoid such conflicts. But since we cannot be sure what will work, we must learn policies for dealing with uncertainty. Which strategies are best to try, and which will avoid the worst mistakes? Thousands and, perhaps, millions of little processes must be involved in how we anticipate, imagine, plan, predict, and prevent—and yet all this proceeds so automatically that we regard it as "ordinary common sense." But if thinking is so complicated, what makes it seem so simple? At first it may seem incredible that our minds could use such intricate machinery and yet be unaware of it.

In general, we're least aware of what our minds do best.

It's mainly when our other systems start to fail that we engage the special agencies involved with what we call "consciousness." Accordingly, we're more aware of simple processes that don't work well than of complex ones that work flawlessly. This means that we cannot trust our offhand judgments about which of the things we do are simple, and which require complicated machinery. Most times, each portion of the mind can only sense how quietly the other portions do their jobs.

2.6 ARE PEOPLE MACHINES?

Many people feel offended when their minds are likened to computer programs or machines. We've seen how a simple tower-building skill can be composed of smaller parts. But could anything like a real mind be made of stuff so trivial?

> "*Ridiculous*," most people say. "*I certainly don't feel like a machine!*"

But if you're not a machine, what makes you an authority on what it feels like to be a machine? A person might reply, "*I think, therefore I know how the mind works.*" But that would be suspiciously like saying, "*I drive my car, therefore I know how its engine works.*" Knowing how to use something is not the same as knowing how it works.

> "*But everyone knows that machines can behave only in lifeless, mechanical ways.*"

This objection seems more reasonable: indeed, a person *ought* to feel offended at being likened to any *trivial* machine. But it seems to me that the word "machine" is getting to be out of date. For centuries, words like "mechanical" made us think of simple devices like pulleys, levers, locomotives, and typewriters. (The word "computerlike" inherited a similar sense of pettiness, of doing dull arithmetic by little steps.) But we ought to recognize that we're still in an early era of machines, with virtually no idea of what they may become. What if some visitor from Mars had come a billion years ago to judge the fate of earthly life from watching clumps of cells that hadn't even learned to crawl? In the same way, we cannot grasp the range of what machines may do in the future from seeing what's on view right now.

Our first intuitions about computers came from experiences with machines of the 1940s, which contained only thousands of parts. But a human brain contains billions of cells, each one complicated by itself and connected to many thousands of others. Present-day computers represent an intermediate degree of complexity; they now have millions of parts, and people already are building billion-part computers for research on Artificial Intelligence. And yet, in spite of what is happening, we continue to use old words as though there had been no change at all. We need to adapt our attitudes to phenomena that work on scales never before conceived. The term "machine" no longer takes us far enough.

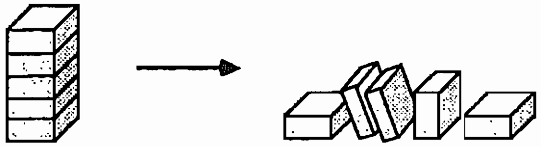
But rhetoric won't settle anything. Let's put these arguments aside and try instead to understand what the vast, unknown mechanisms of the brain may do. Then we'll find more self-respect in knowing what wonderful machines we are.

CHAPTER 3

CONFLICT AND COMPROMISE

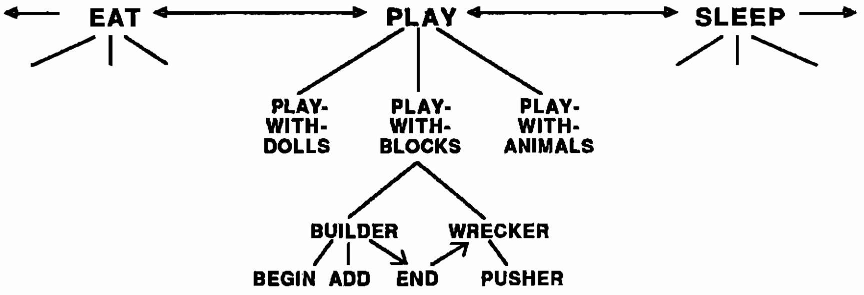
3.1 CONFLICT

Most children not only like to build, they also like to knock things down. So let's imagine another agent called *Wrecker,* whose specialty is knocking-down. Our child loves to hear the complicated noises and watch so many things move all at once.



Suppose *Wrecker* gets aroused, but there's nothing in sight to smash. Then *Wrecker* will have to get some help—by putting *Builder* to work, for example. But what if, at some later time, *Wrecker* considers the tower to be high enough to smash, while *Builder* wants to make it taller still? Who could settle that dispute?

The simplest policy would be to leave that decision to *Wrecker,* who was responsible for activating *Builder* in the first place. But in a more realistic picture of a child's mind, such choices would depend on many other agencies. For example, let's assume that both *Builder* and *Wrecker* were originally activated by a higher-level agent, *Play-with-Blocks.* Then, a conflict might arise if *Builder* and *Wrecker* disagree about whether the tower is high enough.



What aroused *Play-with-Blocks* in the first place? Perhaps some even higher-level agent, *Play,* was active first. Then, inside *Play,* the agent *Play-with-Blocks* achieved control, in spite of two competitors, *Play-with-Dolls* and *Play-with-Animals.* But even *Play* itself, their mutual superior in chief, must have had to compete with other higher-level agencies like *Eat* and *Sleep.* For, after all, a child's play is not an isolated thing but always happens in the context of other real-life concerns. Whatever we may choose to do, there are always other things we'd also like to do.

In several sections of this book, I will assume that conflicts between agents tend to migrate upward to higher levels. For example, any prolonged conflict between *Builder* and *Wrecker* will tend to weaken their mutual superior, *Play-with-Blocks.* In turn, this will reduce *Play-with-Blocks*' ability to suppress *its* rivals, *Play-with-Dolls* and *Play-with-Animals.* Next, if *that* conflict isn't settled soon, it will weaken the agent *Play* at the next-higher level. Then *Eat* or *Sleep* might seize control.

3.2 NONCOMPROMISE

To settle arguments, nations develop legal systems, corporations establish policies, and individuals may argue, fight, or compromise—or turn for help to mediators that lie outside themselves. What happens when there are conflicts inside minds?

Whenever several agents have to compete for the same resources, they are likely to get into conflicts. If those agents were left to themselves, the conflicts might persist indefinitely, and this would leave those agents paralyzed, unable to accomplish any goal. What happens then? We'll assume that those agents' supervisors, too, are under competitive pressure and likely to grow weak themselves whenever their subordinates are slow in achieving their goals, no matter whether because of conflicts between them or because of individual incompetence.

> **The Principle of Noncompromise:** *The longer an internal conflict persists among an agent's subordinates, the weaker becomes that agent's status among its own competitors. If such internal problems aren't settled soon, other agents will take control and the agents formerly involved will be "dismissed."*

So long as playing with blocks goes well, *Play* can maintain its strength and keep control. In the meantime, though, the child may also be growing hungry and sleepy, because other processes are arousing the agents *Eat* and *Sleep*. So long as *Eat* and *Sleep* are not yet strongly activated, *Play* can hold them both at bay. However, any conflict inside *Play* will weaken it and make it easier for *Eat* or *Sleep* to take over. Of course, *Eat* or *Sleep* must conquer in the end, since the longer they wait, the stronger they get.

We see this in our own experience. We all know how easy it is to fight off small distractions when things are going well. But once some trouble starts inside our work, we become increasingly impatient and irritable. Eventually we find it so hard to concentrate that the least disturbance can allow another, different, interest to take control. Now, when any of our agencies loses the power to control what other systems do, that doesn't mean it has to cease its own internal activity. An agency that has lost control can continue to work inside itself—and thus become prepared to seize a later opportunity. However, we're normally unaware of all those other activities proceeding deep inside our minds.

Where does it stop, this process of yielding control to other agencies? Must every mind contain some topmost center of control? Not necessarily. We sometimes settle conflicts by appealing to superiors, but other conflicts never end and never cease to trouble us.

At first, our principle of noncompromise may seem too extreme. After all, good human supervisors plan ahead to avoid conflicts in the first place, and—when they can't—they try to settle quarrels locally before appealing to superiors. But we should not try to find a close analogy between the low-level agents of a single mind and the members of a human community. Those tiny mental agents simply cannot know enough to be able to negotiate with one another or to find effective ways to adjust to each other's interference. Only larger agencies could be resourceful enough to do such things. Inside an actual child, the agencies responsible for *Building* and *Wrecking* might indeed become versatile enough to negotiate by offering support for one another's goals. "*Please, **Wrecker,** wait a moment more till **Builder** adds just one more block: it's worth it for a louder crash!*"

3.3 HIERARCHIES

bu•reauc´ra•cy *n.* the administration of government
through departments and subdivisions managed by
sets of officials following an inflexible routine.

—*Webster's Unabridged Dictionary*

As an agent, *Builder* does no physical work but merely turns on *Begin*, *Add*, and *End*. Similarly, *Add* just orders *Find*, *Put*, and *Get* to do their jobs. Then these divide into agents like *Move* and *Grasp*. It seems that it will never stop—this breaking-down to smaller things. Eventually, it all must end with agents that do actual work, but there are many steps before we get to all the little muscle-motor agents that actually move the arms and hands and finger joints. Thus *Builder* is like a high-level executive, far removed from those subordinates who actually produce the final product.

Does this mean that *Builder*'s administrative work is unimportant? Not at all. Those lower-level agents need to be controlled. It's much the same in human affairs. When any enterprise becomes too complex and large for one person to do, we construct organizations in which certain agents are concerned, not with the final result, but only with what some other agents do. Designing any society, be it human or mechanical, involves decisions like these:

Which agents choose which others to do what jobs?
Who will decide which jobs are done at all?
Who decides what efforts to expend?
How will conflicts be settled?

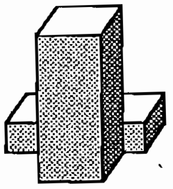
How much of ordinary human thought has *Builder*'s character? The *Builder* we described is not much like a human supervisor. It doesn't decide which agents to assign to which jobs, because that has already been arranged. It doesn't plan its future work but simply carries out fixed steps until *End* says the job is done. Nor has it any repertoire of ways to deal with unexpected accidents.

Because our little mental agents are so limited, we should not try to extend very far the analogy between them and human supervisors and workers. Furthermore, as we'll shortly see, the relations between mental agents are not always strictly hierarchical. And in any case, such roles are always relative. To *Builder*, *Add* is a subordinate, but to *Find*, *Add* is a boss. As for yourself, it all depends on how you live. Which sorts of thoughts concern you most—the orders you are made to take or those you're being forced to give?

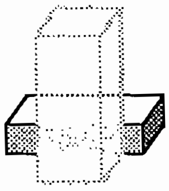
3.4 HETERARCHIES

A hierarchical society is like a tree in which the agent at each branch is exclusively responsible for the agents on the twigs that branch from it. This pattern is found in every field, because dividing work into parts like that is usually the easiest way to start solving a problem. It is easy to construct and understand such organizations because each agent has only a single job to do: it needs only to "look up" for instructions from its supervisor, then "look down" to get help from its subordinates.

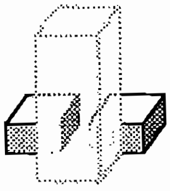
But hierarchies do not always work. Consider that when two agents need to use each other's skills, then neither one can be "on top." Notice what happens, for example, when you ask your vision-system to decide whether the left-side scene below depicts three blocks—or only two.



What you see.



Is it this?



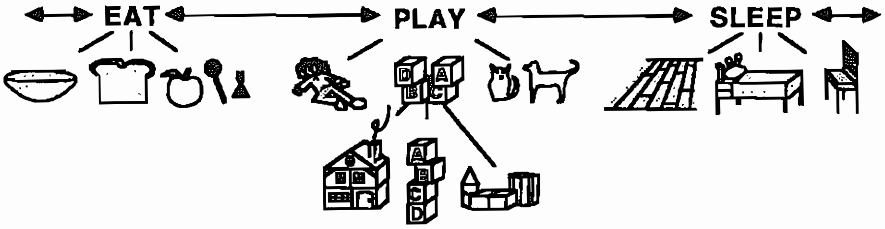
Or this?

The agent *See* could answer that if it could *Move* the front block out of the line of view. But, in the course of doing that, *Move* might have to *See* if there were any obstacles that might interfere with the arm's trajectory. At such a moment, *Move* would be working for *See*, and *See* would be working for *Move*, both at the same time. This would be impossible inside a simple hierarchy.

Most of the diagrams in the early parts of this book depict simple hierarchies. Later, we'll see more cross-connected rings and loops—when we are forced to consider the need for memory, which will become a constant subject of concern in this book. People often think of memory in terms of keeping records of the past, for recollecting things that happened in earlier times. But agencies also need other kinds of memory as well. *See*, for example, requires some sort of temporary memory in order to keep track of what next to do, when it starts one job before its previous job is done. If each of *See*'s agents could do only one thing at a time, it would soon run out of resources and be unable to solve complicated problems. But if we have enough memory, we can arrange our agents into circular loops and thus use the same agents over and over again to do parts of several different jobs at the same time.

3.5 DESTRUCTIVENESS

In any actual child's mind, the urge to *Play* competes with other demanding urges, such as *Eat* and *Sleep*. What happens if another agent wrests control from *Play*, and what happens to the agents *Play* controlled?



Suppose that our child is called away, no matter whether by someone else or by an internal urge like *Sleep*. What happens to the processes remaining active in the mind? One part of the child may still want to play, while another part wants to sleep. Perhaps the child will knock the tower down with a sudden, vengeful kick. What does it mean when children make such scenes? Is it that inner discipline breaks down to cause those savage acts? Not necessarily. Those "childish" acts might still make sense in other ways.

> *Smashing takes so little time that* **Wrecker,** *freed from* **Play**'*s constraint, need persist for only one more kick to gain the satisfaction of a final crash.*
>
> *Though childish violence might seem senseless by itself, it serves to communicate frustration at the loss of goal. Even if the parent scolds, that just confirms how well the message was transmitted and received.*
>
> *Destructive acts can serve constructive goals by leaving fewer problems to be solved. That kick may leave a mess outside, yet tidy up the child's mind.*

When children smash their treasured toys, we shouldn't ask for *the* reason why—since no such act has a single cause. Besides, it isn't true, in a human mind, that when *Sleep* starts, then *Play* must quit and all its agents have to cease. A real child can go to bed—yet still build towers in its head.

3.6 PAIN AND PLEASURE SIMPLIFIED

When you're in pain, it's hard to keep your interest in other things. You feel that nothing's more important than finding some way to stop the pain. That's why pain is so powerful: it makes it hard to think of anything else. Pain simplifies your point of view.

When something gives you pleasure, then, too, it's hard to think of other things. You feel that nothing's more important than finding a way to make that pleasure last. That's why pleasure is so powerful. It also simplifies your point of view.

Pain's power to distract us from our other goals is not an accident; that's how it helps us to survive. Our bodies are endowed with special nerves that detect impending injuries, and the signals from these nerves for pain make us react in special ways. Somehow, they disrupt our concerns with long-term goals—thus forcing us to focus on immediate problems, perhaps by transferring control to our lowest-level agencies. Of course, this can do more harm than good, especially when, in order to remove the source of pain, one has to make a complex plan. Unfortunately, pain interferes with making plans by undermining interest in anything that's not immediate. Too much suffering diminishes us by restricting the complexities that constitute our very selves. It must be the same for pleasure as well.

We think of pleasure and pain as opposites, since pleasure makes us draw its object near while pain impels us to reject its object. We also think of them as similar, since both make rival goals seem small by turning us from other interests. They both distract. Why do we find such similarities between antagonistic things? Sometimes two seeming opposites are merely two extremes along a single scale, or one of them is nothing but the absence of the other—as in the case of sound and silence, light and darkness, interest and unconcern. But what of opposites that are genuinely different, like pain and pleasure, fear and courage, hate and love?

> *In order to appear opposed, two things must serve related goals—or otherwise engage the selfsame agencies.*

Thus, affection and abhorrence both involve our attitudes toward relationships; and pleasure and pain both engage constraints that simplify our mental scenes. The same goes for courage and cowardice: each does best by knowing both. When on attack, you have to press against whatever weakness you can find in your opponent's strategy. When on defense, it's much the same: you still must guess the other's plan.

CHAPTER 4

THE SELF

We are what we pretend to be, so we must be careful about what we pretend to be.

—KURT VONNEGUT

4.1 THE SELF

self *n.* 1. *the identity, character, or essential qualities of any person or thing.* 2. *the identity, personality, individuality, etc. of a given person; one's own person as distinct from all others.*
—*Webster's Unabridged Dictionary*

We all believe that human minds contain those special entities we call selves. But no one agrees about what they are. To keep things straight, I shall write "self" when speaking in a general sense about an entire person and reserve "Self" for talking about that more mysterious sense of personal identity. Here are some of the things people say about the Self:

> *Self is the part of mind that's really me, or rather, it's the part of me—that is, part of my mind—that actually does the thinking and wanting and deciding and enjoying and suffering. It's the part that's most important to me because it's that which stays the same through all experience—the* **identity** *which ties everything together. And whether you can treat it scientifically or not, I know it's there, because it's me. Perhaps it's the sort of thing that Science can't explain.*

This isn't much of a definition, but I don't think it is a good idea to try to find a better one. It often does more harm than good to force definitions on things we don't understand. Besides, only in logic and mathematics do definitions ever capture concepts perfectly. The things we deal with in practical life are usually too complicated to be represented by neat, compact expressions. Especially when it comes to understanding minds, we still know so little that we can't be sure our ideas about psychology are even aimed in the right directions. In any case, one must not mistake defining things for knowing what they are. You can know what a tiger is without defining it. You may define a tiger, yet know scarcely anything about it.

Even if our old ideas about the mind are wrong, we can learn a lot by trying to understand why we believe them. Instead of asking, "*What are Selves?*" we can ask, instead, "*What are our ideas about Selves?*"—and then we can ask, "*What psychological functions do those ideas serve?*" When we do this, it shows us that we do not have one such idea, but many.

Our ideas about our Selves include beliefs about what we *are*. These include beliefs both about what we are capable of doing and about what we may be disposed to do. We exploit these beliefs whenever we solve problems or make plans. I'll refer to them, rather vaguely, as a person's *self-images*. In addition to our self-images, our ideas about ourselves also include ideas about what we'd *like* to be and ideas about what we *ought* to be. These, which I'll call a person's *self-ideals*, influence each person's growth from infancy, but we usually find them hard to express because they're inaccessible to consciousness.

4.2 ONE SELF OR MANY?

One common image of the Self suggests that every mind contains some sort of Voyeur-Puppeteer inside—to feel and want and choose for us the things we feel, want, and choose. But if we had *those* kinds of Selves, what would be the use of having Minds? And, on the other hand, if Minds could do such things themselves, why have Selves? Is this concept of a Self of any real use at all? It is indeed—provided that we think of it not as a centralized and all-powerful entity, but as a society of ideas that include both our images of what the mind is and our ideals about what it ought to be.

Besides, we're often of two minds about ourselves. Sometimes we regard ourselves as single, self-coherent entities. Other times we feel decentralized or dispersed, as though we were made of many different parts with different tendencies. Contrast these views:

> **SINGLE-SELF VIEW.** *"I think, I want, I feel. It's me, myself, who thinks my thoughts. It's not some nameless crowd or cloud of selfless parts."*
>
> **MULTIPLE-SELF VIEW.** *"One part of me wants this, another part wants that. I must get better control of myself."*

We're never wholly satisfied with either view. We all sense feelings of disunity, conflicting motives, compulsions, internal tensions, and dissensions. We carry on negotiations in our head. We hear scary tales in which some person's mind becomes enslaved by compulsions and commands that seem to come from somewhere else. And the times we feel most reasonably unified can be just the times that others see us as the most confused.

But if there is no single, central, ruling Self inside the mind, what makes us feel so sure that one exists? What gives that myth its force and strength? A paradox: perhaps it's *because* there are no persons in our heads to make us do the things we want—nor even ones to make us *want to want*—that we construct the myth that *we're* inside ourselves.

4.3 THE SOUL

And we thank Thee that darkness reminds us of light.
—T. S. Eliot

A common concept of the soul is that the essence of a self lies in some spark of invisible light, a thing that cowers out of body, out of mind, and out of sight. But what might such a symbol mean? It carries a sense of anti-self-respect: that there is no significance in anyone's accomplishments.

People ask if machines can have souls. And I ask back whether souls can learn. It does not seem a fair exchange—if souls can live for endless time and yet not use that time to learn—to trade all change for changelessness. And that's exactly what we get with inborn souls that cannot grow: a destiny the same as death, an ending in a permanence incapable of any change and, hence, devoid of intellect.

Why try to frame the value of a Self in such a singularly frozen form? The art of a great painting is not in any one idea, nor in a multitude of separate tricks for placing all those pigment spots, but in the great network of relationships among its parts. Similarly, the agents, raw, that make our minds are by themselves as valueless as aimless, scattered daubs of paint. What counts is what we make of them.

We all know how an ugly husk can hide an unexpected gift, like a treasure buried in the dust or a graceless oyster bearing a pearl. But minds are just the opposite. We start as little embryos, which then build great and wondrous selves—whose merit lies entirely within their own coherency. The value of a human self lies not in some small, precious core, but in its vast, constructed crust.

What are those old and fierce beliefs in spirits, souls, and essences? *They're all insinuations that we're helpless to improve ourselves.* To look for our virtues in such thoughts seems just as wrongly aimed a search as seeking art in canvas cloths by scraping off the painter's works.

4.4 THE CONSERVATIVE SELF

How do we control our minds? Ideally, we first choose what we want to do, then make ourselves do it. But that's harder than it sounds: we spend our lives in search of schemes for self-control. We celebrate when we succeed, and when we fail, we're angry with ourselves for not behaving as we wanted to—and then we try to scold or shame or bribe ourselves to change our ways. But wait! How could a self be angry with itself? Who would be mad at whom? Consider an example from everyday life.

> *I was trying to concentrate on a certain problem but was getting bored and sleepy. Then I imagined that one of my competitors, Professor Challenger, was about to solve the same problem. An angry wish to frustrate Challenger then kept me working on the problem for a while. The strange thing was, this problem was not of the sort that ever interested Challenger.*

What makes us use such roundabout techniques to influence ourselves? Why be so indirect, inventing misrepresentations, fantasies, and outright lies? Why can't we simply tell ourselves to do the things we want to do?

To understand how something works, one has to know its purposes. Once, no one understood the heart. But as soon as it was seen that hearts move blood, a lot of other things made sense: those things that looked like pipes and valves were really pipes and valves indeed—and anxious, pounding, pulsing hearts were recognized as simple pumps. New speculations could then be formed: was this to give our tissues drink or food? Was it to keep our bodies warm or cool? For sending messages from place to place? In fact, all those hypotheses were correct, and when that surge of functional ideas led to the guess that blood can carry air as well, more puzzle parts fell into place.

To understand what we call the Self, we first must see what Selves are for. *One function of the Self is to keep us from changing too rapidly.* Each person must make some long-range plans in order to balance single-purposeness against attempts to do everything at once. But it is not enough simply to instruct an agency to start to carry out our plans. We also have to find some ways to constrain the changes we might later make—to prevent ourselves from turning those plan-agents off again! If we changed our minds too recklessly, we could never know what we might want next. We'd never get much done because we could never depend on ourselves.

Those ordinary views are wrong that hold that Selves are magic, self-indulgent luxuries that enable our minds to break the bonds of natural cause and law. Instead, those Selves are practical necessities. The myths that say that Selves embody special kinds of liberty are merely masquerades. Part of their function is to hide from us the nature of our self-ideals—the chains we forge to keep ourselves from wrecking all the plans we make.

4.5 EXPLOITATION

Let's look more closely at that episode of Professor Challenger. Apparently, what happened was that my agency for *Work* exploited *Anger* to stop *Sleep*. But why should *Work* use such a devious trick?

To see why we have to be so indirect, consider some alternatives. If *Work* could simply turn off *Sleep*, we'd quickly wear our bodies out. If *Work* could simply switch *Anger* on, we'd be fighting all the time. Directness is too dangerous. We'd die.

Extinction would be swift indeed for species that could simply switch off hunger or pain. Instead, there must be checks and balances. We'd never get through one full day if any agency could seize and hold control over all the rest. This must be why our agencies, in order to exploit each other's skills, have to discover such roundabout pathways. All direct connections must have been removed in the course of our evolution.

This must be one reason why we use fantasies: to provide the missing paths. You may not be able to make yourself angry simply by deciding to be angry, but you can still imagine objects or situations that *make* you angry. In the scenario about Professor Challenger, my agency *Work* exploited a particular memory to arouse my *Anger*'s tendency to counter *Sleep*. This is typical of the tricks we use for self-control.

Most of our self-control methods proceed unconsciously, but we sometimes resort to conscious schemes in which we offer rewards to ourselves: "*If I can get this project done, I'll have more time for other things.*" However, it is not such a simple thing to be able to bribe yourself. To do it successfully, you have to discover which mental incentives will actually work on yourself. This means that you—or rather, your agencies—have to learn something about one another's dispositions. In this respect the schemes we use to influence ourselves don't seem to differ much from those we use to exploit other people—and, similarly, they often fail. When we try to induce ourselves to work by offering ourselves rewards, we don't always keep our bargains; we then proceed to raise the price or even to deceive ourselves, much as one person may try to conceal an unattractive aspect of a bargain from another person.

Human self-control is no simple skill, but an ever-growing world of expertise that reaches into everything we do. Why is it that, in the end, so few of our self-incentive tricks work well? Because, as we have seen, directness is too dangerous. If self-control were easy to obtain, we'd end up accomplishing nothing at all.

4.6 SELF-CONTROL

Those who really seek the path to Enlightenment dictate terms to their mind. Then they proceed with strong determination.
—Buddha

The episode of Professor Challenger showed just one way we can control ourselves: by exploiting an emotional aversion in order to accomplish an intellectual purpose. Consider all the other kinds of tricks we use to try to force ourselves to work when we're tired or distracted.

WILLPOWER: Tell yourself, "Don't give in to that," or, "Keep on trying."

Such self-injunctions can work at first—but finally they always fail, as though some engine in the mind runs out of fuel. Another style of self-control involves more physical activity:

ACTIVITY: Move around. Exercise. Inhale. Shout.

Certain physical acts are peculiarly effective, especially the facial expressions involved in social communication: they affect the sender as much as the recipient.

EXPRESSION: Set jaw. Stiffen upper lip. Furrow brow.

Another kind of stimulating act is moving to a stimulating place. And we often perform actions that directly change the brain's chemical environment.

CHEMISTRY: Take coffee, amphetamines, or other brain-affecting drugs.

Then there are actions in the mind with which we set up thoughts and fantasies that move our own emotions, arousing hopes and fears through self-directed offers, bribes, and even threats.

EMOTION: "If I win, there's much to gain, but more to lose if I fail!"

Perhaps most powerful of all are those actions that promise gain or loss of the regard of certain special persons.

ATTACHMENT: Imagine admiration if you succeed—or disapproval if you fail—especially from those to whom you are attached.

So many schemes for self-control! How do we choose which ones to use? There isn't any easy way. Self-discipline takes years to learn; it grows inside us stage by stage.

4.7 LONG-RANGE PLANS

In the search for truth there are certain questions that are not important. Of what material is the universe constructed? Is the universe eternal? Are there limits or not to the universe? What is the ideal form of organization for human society? If a man were to postpone his search and practice for Enlightenment until such questions were solved, he would die before he found the path.
—BUDDHA

We often become involved in projects that we can't complete. It is easy to solve small problems because we can treat them as though they were detached from all our other goals. But it is different for projects that span larger portions of our lives, like learning a trade, raising a child, or writing a book. We cannot simply "decide" or "choose" to accomplish an enterprise that makes a large demand for time, because it will inevitably conflict with other interests and ambitions. Then we'll be forced to ask questions like these:

What must I give up for this?
What will I learn from it?
Will it bring power and influence?
Will I remain interested in it?
Will other people help me with it?
Will they still like me?

Perhaps the most difficult question of all is, "*How will adopting this goal change me?*" Just wanting to own a large, expensive house, for instance, can lead to elaborate thoughts like these:

> *"That means I'd have to save for years and not get other things I'd like. I doubt that I could bear it. True, I could reform myself, and try to be more thrifty and deliberate. But that's just not the sort of person I am."*

Until such doubts are set aside, all the plans we make will be subject to the danger that we may "change our mind." So how can any long-range plan succeed? The easiest path to "self-control" is doing only what one is already disposed to do.

Many of the schemes we use for self-control are the same as those we learn to use for influencing other people. We make ourselves behave by exploiting our own fears and desires, offering ourselves rewards, or threatening the loss of what we love. But when short-range tricks won't keep us to our projects for long enough, we may need some way to make changes that won't let us change ourselves back again. I suspect that, in order to commit ourselves to our largest, most ambitious plans, we learn to exploit agencies that operate on larger spans of time.

Which are our slowest-changing agencies of all? Later we'll see that these must include the silent, hidden agencies that shape what we call *character*. These are the systems that are concerned not merely with the things we *want*, but with what we *want ourselves to be*—that is, the ideals we set for ourselves.

4.8 IDEALS

We usually reserve the word "ideals" to refer to how we think we ought to conduct our ethical affairs. But I'll use the term in a broader sense, to include the standards we maintain—consciously or otherwise—for how we ought to think about ordinary matters.

We're always involved with goals of varying spans and scales. What happens when a transient inclination clashes with a long-term self-ideal? What happens, for that matter, when our ideals disagree among themselves, as when there is an inconsistency between the things we want to do and those we feel we ought to do? These disparities give rise to feelings of discomfort, guilt, and shame. To lessen such disturbances, we must either change the things we do—or change the ways we feel. Which should we try to modify—our immediate wants or our ideals? Such conflicts must be settled by the multilayered agencies that are formed in the early years of the growth of our personalities.

In childhood, our agencies acquire various types of goals. Then we grow in overlapping waves, in which our older agencies affect the making of the new. This way, the older agencies can influence how our later ones will behave. Outside the individual, similar processes go on in every human community; we find children "taking after" persons other than themselves by absorbing values from their parents, families, and peers, even from the heroes and villains of mythology.

Without enduring self-ideals, our lives would lack coherence. As individuals, we'd never be able to trust ourselves to carry out our personal plans. In a social group, no one person would be able to trust the others. A working society must evolve mechanisms that stabilize ideals—and many of the social principles that each of us regards as personal are really "long-term memories" in which our cultures store what they have learned across the centuries.

CHAPTER 5

INDIVIDUALITY

PUNCH AND JUDY, TO THEIR AUDIENCE

Our puppet strings are hard to see,
So we perceive ourselves as free,
Convinced that no mere objects could
Behave in terms of bad and good.

To you, we mannikins seem less
than live, because our consciousness
is that of dummies, made to sit
on laps of gods and mouth their wit;

Are you, our transcendental gods,
likewise dangled from your rods,
and need, to show spontaneous charm,
some higher god's inserted arm?

We seem to form a nested set,
with each the next one's marionette,
who, if you asked him, would insist
that he's the last ventriloquist.

—THEODORE MELNECHUK

5.1 CIRCULAR CAUSALITY

Whenever we can, we like to explain things in terms of simple cause and effect. We explained the case of Professor Challenger by assuming that my wish to *Work* came first, then *Work* exploited *Anger*'s aptitude for fighting *Sleep*. But in real life the causal relations between feelings and thoughts are rarely so simple. My desire to work and my annoyance with Challenger were probably so intermingled, all along, that it is inappropriate to ask which came first, *Anger* or *Work*. Most likely, *both* agencies exploited one another simultaneously, thus combining both into a single fiendish synthesis that accomplished two goals at once; *Work* thus got to do its work —and, thereby, injured Challenger! (In an academic rivalry, a technical accomplishment can hurt more than a fist.) Two goals can support each other.

A causes B *"John wanted to go home because he felt tired of work."*
B causes A *"John felt tired of work because he wanted to go home."*

There need be no "first cause" since John could start out with both distaste for work and inclination to go home. Then a loop of circular causality ensues, in which each goal gains support from the other until their combined urge becomes irresistible. We're always enmeshed in causal loops. Suppose you had borrowed past your means and later had to borrow more in order to pay the interest on your loan. If you were asked what the difficulty was, it would not be enough to say simply, *"Because I have to pay the interest,"* or to say only, *"Because I have to pay the principal."* Neither alone is the actual cause, and you'd have to explain that you're caught in a loop.

We often speak of "straightening things out" when we're involved in situations that seem too complicated. It seems to me that this metaphor reflects how hard it is to find one's way through a maze that has complicated loops in it. In such a situation, we always try to find a "path" through it by seeking "causal" explanations that go in only one direction. There's a good reason for doing this.

> *There are countless different types of networks that contain loops. But all networks that contain no loops are basically the same: each has the form of a simple chain.*

Because of this, we can apply the very same types of reasoning to *everything* we can represent in terms of chains of causes and effects. Whenever we accomplish that, we can proceed from start to end without any need for a novel thought; that's what we mean by "straightening out." But frequently, to construct such a path, we have to ignore important interactions and dependencies that run in other directions.

5.2 UNANSWERABLE QUESTIONS

And while it shall please thee to continue me in this world, where there is much to be done and little to be known, teach me, by thy Holy Spirit, to withdraw my mind from unprofitable and dangerous enquiries, from difficulties vainly curious, and doubts impossible to be solved.

—SAMUEL JOHNSON

When we reflect on anything for long enough, we're likely to end up with what we sometimes call "basic" questions—ones we can see no way at all to answer. For we have no perfect way to answer even this question: *How can one tell when a question has been properly answered?*

What caused the universe, and why?	*What is the purpose of life?*
How can you tell which beliefs are true?	*How can you tell what is good?*

These questions seem different on the surface, but all of them share one quality that makes them impossible to answer: *all of them are circular!* You can never find a final cause, since you must always ask one question more: "*What caused that cause?*" You can never find any ultimate goal, since you're always obliged to ask, "*Then what purpose does **that** serve?*" Whenever you find out why something is good—or is true—you still have to ask what makes *that* reason good and true. No matter what you discover, at every step, these kinds of questions will always remain, because you have to challenge every answer with, "*Why should I accept **that** answer?*" Such circularities can only waste our time by forcing us to repeat, over and over and over again, "*What good is Good?*" and, "*What god made God?*"

When children keep on asking, "*Why?*" we adults learn to deal with this by simply saying, "*Just because!*" This may seem obstinate, but it's also a form of self-control. What stops adults from dwelling on such questions endlessly? The answer is that every culture finds special ways to deal with these questions. One way is to brand them with shame and taboo; another way is to cloak them in awe or mystery; both methods make those questions undiscussable. Consensus is the simplest way—as with those social styles and trends wherein we each accept as true whatever all the others do. I think I once heard W. H. Auden say, "*We are all here on earth to help others. What I can't figure out is what the others are here for.*"

All human cultures evolve institutions of law, religion, and philosophy, and these institutions both adopt specific answers to circular questions and establish authority-schemes to indoctrinate people with those beliefs. One might complain that such establishments substitute dogma for reason and truth. But in exchange, they spare whole populations from wasting time in fruitless reason loops. Minds can lead more productive lives when working on problems that can be solved.

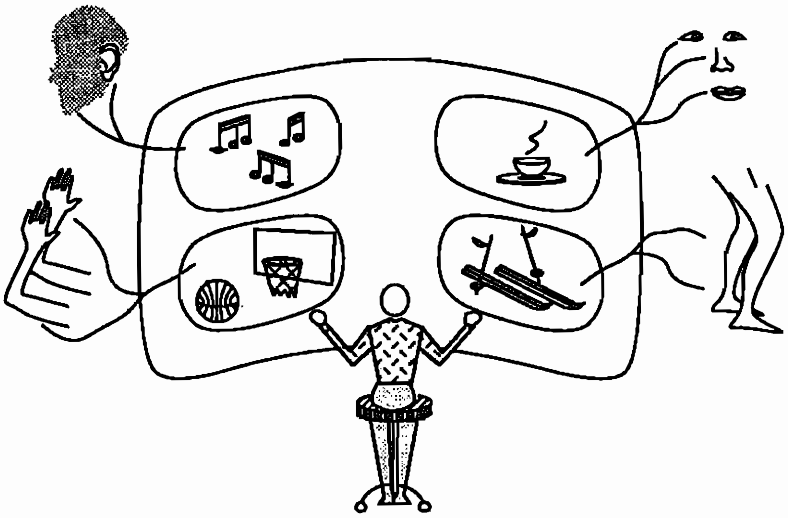
But when thinking keeps returning to its source, it doesn't always mean something's wrong. For circular thinking can lead to growth when it results, at each return, in deeper and more powerful ideas. Then, because we can communicate, such systems of ideas may even find the means to cross the boundaries of selfish selves—and thus take root in other minds. This way, a language, science, or philosophy can transcend the limitation of each single mind's mortality. Now, we cannot know that any individual is destined for some paradise. Yet certain religions are oddly right; they manage to achieve their goal of offering an afterlife—if only to their own strange souls.

5.3 THE REMOTE-CONTROL SELF

When people have no answers to important questions, they often give some anyway.

What controls the brain?	*The Mind.*
What controls the mind?	*The Self.*
What controls the Self?	*Itself.*

To help us think about how our minds are connected to the outer world, our culture teaches schemes like this:



This diagram depicts our sensory machinery as sending information to the brain, wherein it is projected on some inner mental movie screen. Then, inside that ghostly theater, a lurking Self observes the scene and then considers what to do. Finally, that Self may act—somehow reversing all those steps—to influence the real world by sending various signals back through yet another family of remote-control accessories.

This concept simply doesn't work. It cannot help for you to think that inside yourself lies someone else who does your work. This notion of "homunculus"—a little person inside each self—leads only to a paradox since, then, *that inner Self requires yet another movie screen inside itself, on which to project what it has seen!* And then, to watch that play-within-a-play, we'd need yet another Self-inside-a-Self—to do the thinking for the last. And then this would all repeat again, as each new Self requires yet another one to do its job!

> *The idea of a single, central Self doesn't explain anything. This is because a thing with no parts provides nothing that we can use as pieces of explanation!*

Then why do we so often embrace the strange idea that what we do is done by Someone Else —that is, our Self? Because so much of what our minds do is hidden from the parts of us that are involved with verbal consciousness.

5.4 PERSONAL IDENTITY

Whate'er the passion—knowledge, fame, or pelf,
Not one will change his neighbor with himself.
—ALEXANDER POPE

Why do we accept that paradoxical image of a central Self inside the self? Because it serves us well in many spheres of practical life. Here are some reasons to regard a person as a single thing.

> **The Physical World:** *Our bodies act like other objects that take up space. Because of that, we must base our plans and decisions on having a single body. Two people cannot fit where there is room for only one—nor can a person walk through walls or stay aloft without support.*
>
> **Personal Privacy:** *When Mary tells Jack something, she must remember to "whom" it was told, and she must not assume that every other person knows it, too. Also, without the concept of an individual, we could have no sense of responsibility.*
>
> **Mental Activity:** *We often find it hard to think two different thoughts at once, particularly when they're similar, because we get "confused" when the same agencies are asked to do different jobs at the same time.*

Why do our mental processes so often seem to us to flow in "streams of consciousness"? Perhaps because, in order to keep control, we have to simplify how we represent what's happening. Then, when that complicated mental scene is "straightened out," it seems as though a single pipeline of ideas were flowing through the mind.

These are all compelling reasons why it helps to see ourselves as singletons. Still, each of us must also learn not only that different people have their own identities, but that the same person can entertain different beliefs, plans, and dispositions at the same time. For finding good ideas about psychology, the single-agent image has become a grave impediment. To comprehend the human mind is surely one of the hardest tasks any mind can face. The legend of the single Self can only divert us from the target of that inquiry.

5.5 FASHION AND STYLE

The notes I handle no better than many pianists. But the pauses between the notes—ah, that is where the art resides!
—ARTUR SCHNABEL

Why do we like so many things that seem to us to have no earthly use? We often speak of this with mixtures of defensiveness and pride.

"Art for Art's sake."
"I find it aesthetically pleasing."
"I just like it."
"There's no accounting for it."

Why do we take refuge in such vague, defiant declarations? "There's no accounting for it" sounds like a guilty child who's been told to keep accounts. And "I just like it" sounds like a person who is hiding reasons too unworthy to admit. However, we often do have sound practical reasons for making choices that have no reasons by themselves but have effects on larger scales.

> **Recognizability:** *The legs of a chair work equally well if made square or round. Then why do we tend to choose our furniture according to systematic styles or fashions? Because familiar styles make it easier for us to recognize and classify the things we see.*
>
> **Uniformity:** *If every object in a room were interesting in itself, our furniture might occupy our minds too much. By adopting uniform styles, we protect ourselves from distractions.*
>
> **Predictability:** *It makes no difference whether a single car drives on the left or on the right. But it makes all the difference when there are many cars! Societies need rules that make no sense for individuals.*

It can save a lot of mental work if one makes each arbitrary choice the way one did before. The more difficult the decision, the more this policy can save. The following observation by my associate, Edward Fredkin, seems important enough to deserve a name:

> **Fredkin's Paradox:** *The more equally attractive two alternatives seem, the harder it can be to choose between them—no matter that, to the same degree, the choice can only matter less.*

No wonder we often can't account for "taste"—if it depends on hidden rules that we use when ordinary reasons cancel out! I do not mean to say that fashion, style, and art are all the same—only that they often share this strategy of using forms that lie beneath the surface of our thoughts. When should we quit reasoning and take recourse in rules of style? Only when we're fairly sure that further thought will just waste time. Perhaps that's why we often feel such a sense of being free from practicality when we make "aesthetic" choices. Such decisions might seem more constrained if we were aware of how they're made. And what about those fleeting hints of guilt we sometimes feel for "just liking" art? Perhaps they're how our minds remind themselves not to abandon thought too recklessly.

5.6 TRAITS

Isn't it remarkable that words can portray human individuals? You might suppose this should be impossible, considering how much there is to say. Then what permits a writer to depict such seemingly real personalities? It is because we all agree on so many things that are left unsaid. For example, we assume that all the characters are possessed of what we call "commonsense knowledge," and we also agree on many generalities about what we call "human nature."

Hostility evokes defensiveness. Frustration arouses aggression.

We also recognize that individuals have particular qualities and traits of character.

Jane is tidy. Mary's timid. Grace is smart.
That's not the sort of thing Charles does. It's not his style.

Why should traits like these exist? Humanists are prone to boast about how hard it is to grasp the measure of a mind. But let's ask instead, "*What makes personalities so* easy *to portray?*" Why, for example, should any person tend toward a general quality of being neat, rather than simply being tidy about some things and messy about others? Why should our personalities show such coherencies? How could it be that a system assembled from a million agencies can be described by short and simple strings of words? Here are some possible reasons.

Selectivity: *First we should face the fact that our images of other minds are often falsely clear. We tend to think of another person's "personality" in terms of that which we can describe—and tend to set aside the rest as though it simply weren't there.*

Style: *To escape the effort of making decisions we consider unimportant, we tend to develop policies that become so systematic that they can be discerned from the outside and characterized as personal traits.*

Predictability: *Because it is hard to maintain friendship without trust, we try to conform to the expectations of our friends. Then, to the extent that we frame our images of our associates in terms of traits, we find ourselves teaching ourselves to behave in accord with those same descriptions.*

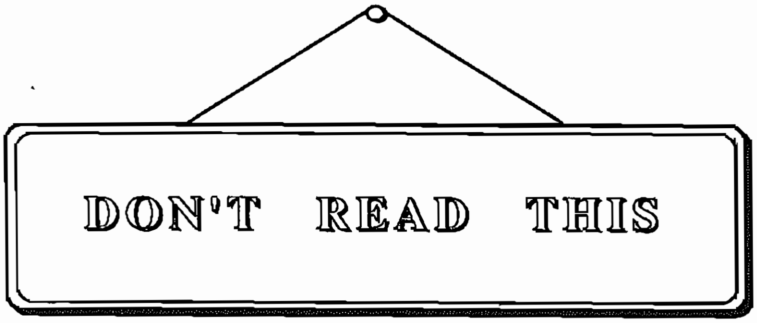
Self-Reliance: *Thus, over time, imagined traits can make themselves actual! For even to carry out our own plans, we must be able to predict what we ourselves are likely to do—and that will become easier the more we simplify ourselves.*

It's nice to be able to trust our friends, but we need to be able to trust ourselves. How can that be possible when we can't be sure what's in our own heads? One way to accomplish this is by thinking of ourselves in terms of traits—and then proceeding to train ourselves to behave according to those self-images. Still, a personality is merely the surface of a person. What we call traits are only the regularities we manage to perceive. We never really know ourselves because there are so many other processes and policies that never show themselves directly in our behavior but work behind the scenes.

5.7 PERMANENT IDENTITY

There are causes for all human suffering, and there is a way by which they may be ended, because everything in the world is the result of a vast concurrence of causes and conditions, and everything disappears as these causes and conditions change and pass away.
—BUDDHA

What do we signify by words like "me," "myself," and "I"? What does a story mean that starts with "In *my* childhood"? What is that strange possession "you," which stays the same throughout your life? Are you the same person you were before you learned to read? You scarcely can imagine, now, how words looked then. Just try to look at these words without reading them:



So far as consciousness is concerned, we find it almost impossible to separate the appearances of things from what they've come to mean to us. But if we cannot recollect how things appeared to us before we learned to link new meanings to those things, what makes us think we can recollect how we ourselves appeared to us in previous times? What would you say if someone asked questions like these:

"Are you the same person now that you once were, before you learned to talk?"
"Of course I am. Why, who else could I be?"
"Do you mean that you haven't changed at all?"
"Of course not. I only mean I'm the same person—the same in some ways, different in others—but still the same me."
"But how can you be the same as the person you were before you had even learned to remember things? Can you even imagine what that was like?"
"Perhaps I can't—yet still there must have been some continuity. Even if I can't remember it, I surely was that person, too."

We all experience that sense of changelessness in spite of change, not only for the past but also for the future, too! Consider how you are generous to future self at present self's expense. Today, you put some money in the bank in order that sometime later you can take it out. Whenever did that future self do anything so good for you? Is "you" the body of those memories whose meanings change only slowly? Is it the never-ending side effects of all your previous experience? Or is it just whichever of your agents change the least as time and life proceed?