OTHER BOOKS BY MARVIN MINSKY:

COMPUTATION: FINITE AND INFINITE MACHINES
SEMANTIC INFORMATION PROCESSING
PERCEPTRONS (WITH SEYMOUR PAPERT)
ROBOTICS

9780671657130TEXT_POD.indd 2 3/10/23 11:26 AM

THE SOCIETY OF MIND

Marvin Minsky

Illustrations by Juliana Lee

SIMON & SCHUSTER PAPERBACKS
NEW YORK • LONDON • TORONTO • SYDNEY

SIMON & SCHUSTER PAPERBACKS
Rockefeller Center
1230 Avenue of the Americas
New York, NY 10020

First paperback edition, 1988

For information about special discounts for bulk purchases,
please contact Simon & Schuster Special Sales:
1-800-456-6798 or business@simonandschuster.com.

Designed by Irving Perkins Associates
Manufactured in the United States of America

30 31 32 33 Pbk.

Library of Congress Cataloging-in-Publication Data

Minsky, Marvin Lee, date.
The society of mind.

Includes index.
1. Intellect. 2. Human information processing.
3. Science—Philosophy. I. Title.
BF431.M553 1986 153 86-20322
ISBN-13: 978-0-671-60740-1
ISBN-10: 0-671-60740-5
ISBN-13: 978-0-671-65713-0 (Pbk)
ISBN-10: 0-671-65713-5 (Pbk)

The author is grateful for permission to reprint the following material:

Excerpts by Theodore Melnechuk. Reprinted by permission of the author.

Excerpts by Manfred Clynes. Reprinted by permission of the author.

From “Burnt Norton” in *Four Quartets* by T. S. Eliot, © 1943 by T. S. Eliot; renewed 1971 by Esme Valerie Eliot. Reprinted by permission of Harcourt Brace Jovanovich, Inc.

From “Choruses from ‘The Rock’ ” in *Collected Poems 1909–1962* by T. S. Eliot. © 1936 by Harcour Brace Jovanovich, Inc.

CONTENTS

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.

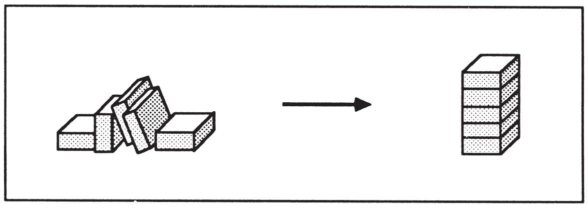
We're always doing several things at once, like planning and walking and talking, and this all seems so natural that we take it for granted. But these processes actually involve more machinery than anyone can understand all at once. So, in the next few sections of this book, we'll focus on just one ordinary 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.

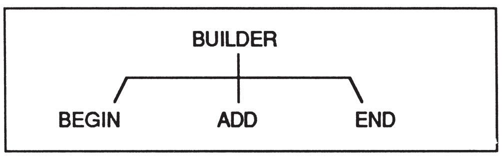
> *In science, one can learn the most by studying what seems the least.*

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.

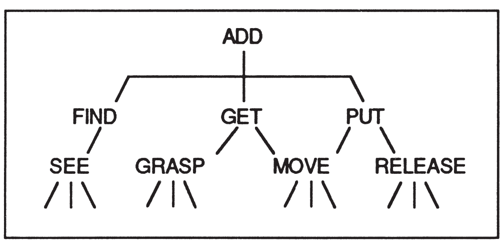


Our child likes to watch a tower grow as each new block is placed on top. But building a tower is too complicated a job for any single, simple agent, so *Builder* has to ask for help from several other agents:



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.*

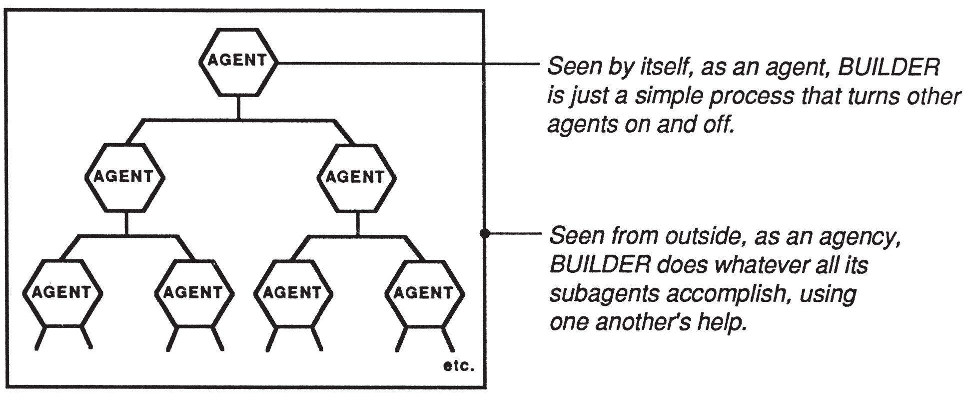
If common sense is so diverse and intricate, what makes it seem so obvious and natural? This illusion of simplicity comes from losing touch with what happened during infancy, when we formed our first abilities. As each new group of skills matures, we build more layers on top of them. As time goes on, 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.



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

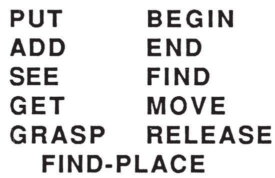
It is the nature of the mind that makes individuals kin, and the differences in the shape, form, or manner of the material atoms out of whose intricate relationships that mind is built are altogether trivial.

—Isaac Asimov

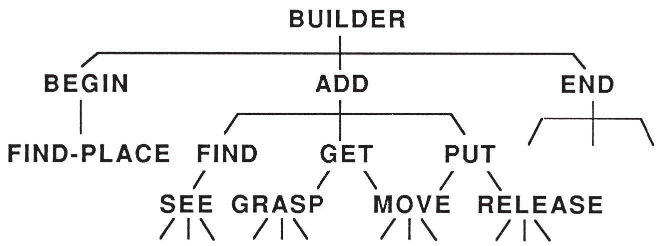
2.1 COMPONENTS AND CONNECTIONS

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.



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 'mousetightness' 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.

9780671657130TEXT_POD.indd 25 3/10/23 11:26 AM

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.

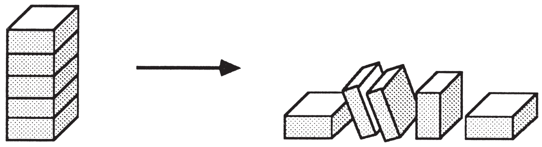
9780671657130TEXT_POD.indd 26 3/10/23 11:26 AM

CHAPTER 3

CONFLICT AND COMPROMISE

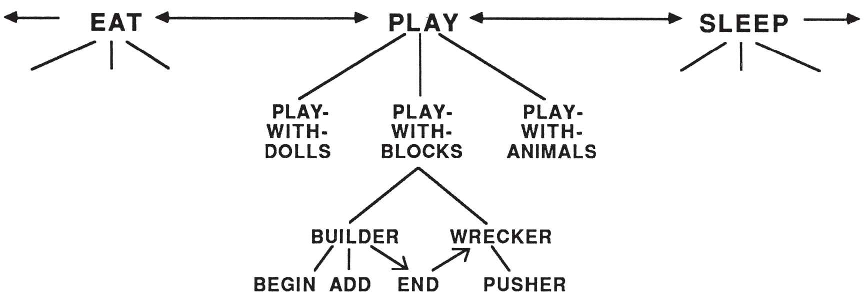
3.1 CONFLICT

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?

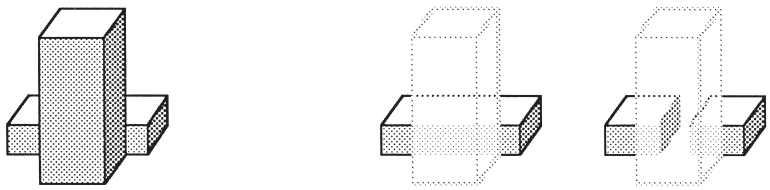
How much of ordinary human thought has *Builder*'s character? The *Builder* we described is not much like a human supervisor. It doesn't decide which agents to assign to which jobs, because that has already been arranged. It doesn't plan its future work but simply carries out fixed steps until *End* 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.



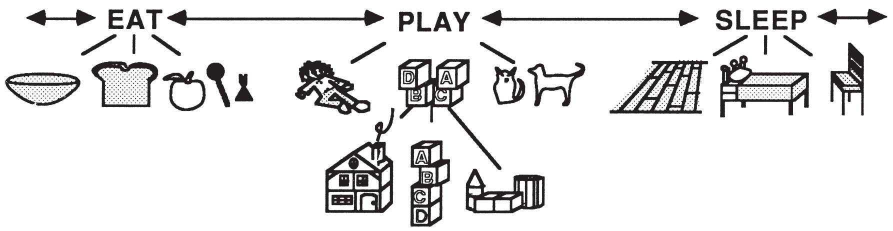
What you see. *Is it this?* *Or this?*

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.

9780671657130TEXT_POD.indd 33 3/10/23 11:26 AM