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.

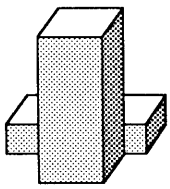
What can we do when things are hard to describe? We start by sketching out the roughest shapes to serve as scaffolds for the rest; it doesn't matter very much if some of those forms turn out partially wrong. Next, draw details to give these skeletons more lifelike flesh. Last, in the final filling-in, 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.

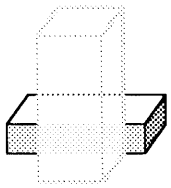
3.4 HETERARCHIES

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.

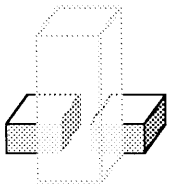
But hierarchies do not always work. Consider that when two agents need to use each other's skills, then neither one can be "on top." Notice what happens, for example, when you ask your vision-system to decide whether the left-side scene below depicts three blocks—or only two.



What you see.



Is it this?



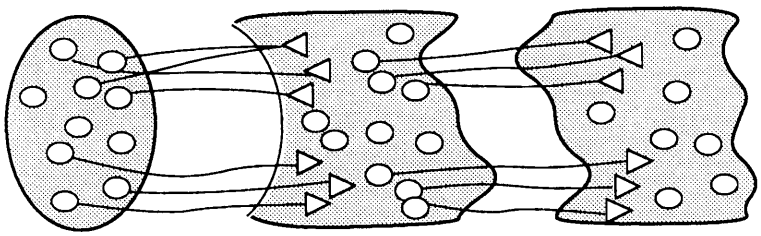
Or this?

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.

6.4 *B*-BRAINS

There *is* one way for a mind to watch itself and still keep track of what's happening. Divide the brain into two parts, A and B. Connect the A-brain's inputs and outputs to the real world—so it can sense what happens there. But don't connect the B-brain to the outer world at all; instead, connect it so that the A-brain is the B-brain's world!



Now A can see and act upon what happens in the outside world—while B can "see" and influence what happens inside A. What uses could there be for such a B? Here are some A-activities that B might learn to recognize and influence.

A seems disordered and confused.	*Inhibit that activity.*
A appears to be repeating itself.	*Make A stop. Do something else.*
A does something B considers good.	*Make A remember this.*
A is occupied with too much detail.	*Make A take a higher-level view.*
A is not being specific enough.	*Focus A on lower-level details.*

This two-part arrangement could be a step toward having a more "reflective" mind-society. The B-brain could do experiments with the A-brain, just as the A-brain can experiment with the body or with the objects and people surrounding it. And just as A can attempt to predict and control what happens in the outer world, B can try to predict and control what A will do. For example, the B-brain could supervise how the A-brain learns, either by making changes in A directly or by influencing A's own learning processes.

Even though B may have no concept of what A's activities mean in relation to the outer world, it is still possible for B to be useful to A. This is because a B-brain could learn to play a role somewhat like that of a counselor, psychologist, or management consultant, who can assess a client's mental strategy without having to understand all the details of that client's profession. Without having any idea of what A's goals are, B might be able to learn to tell when A is not accomplishing them but only going around in circles or wandering, confused because certain A-agents are repeating the same things over and over again. Then B might try some simple remedies, like suppressing some of those A-agents. To be sure, this could also result in B's activities becoming nuisances to A. For example, if A had the goal of adding up a long column of numbers, B might start to interfere with this because, from B's point of view, A appears to have become trapped in a repetitive loop. This could cause a person accustomed to more variety to find it difficult to concentrate on such a task and complain of being bored.

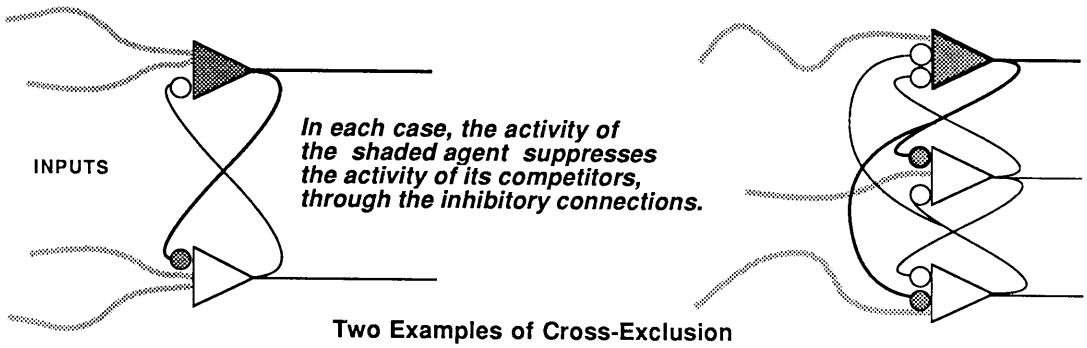
To the extent that the B-brain knows what is happening in A, the entire system could be considered to be partly "self-aware." However, if we connect A and B to "watch" each other too closely, then anything could happen, and the entire system might become unstable. In any case, there is no reason to stop with only two levels; we could connect a C-brain to watch the B-brain, and so on.

16.4 CROSS-EXCLUSION

An ordinary single-bodied animal can only move in one direction at a time, and this tends to constrain it to work toward only one goal at a time. For example, when such an animal needs water urgently, its specialist for "thirst" takes control; however, if cold is paramount, finding warmth takes precedence. But if several urgent needs occur at once, there must be a way to select one of them. One scheme for this might use some sort of central marketplace, in which the urgencies of different goals compete and the highest bidder takes control. However, that strategy is prone to fall into a funny, fatal indecisiveness. To see the problem, imagine that our animal is both very hungry and very thirsty.

> *Suppose that our animal's hunger is, at first, just slightly more urgent than its thirst. So it sets out on a trek toward the North Plain, where food is usually found. When it arrives and takes a bite of food, its thirst instantly takes precedence over its need for food!*
>
> *Now that thirst has top priority, our animal sets out on the long journey toward South Lake. But once it arrives and takes one satisfying sip, the balance instantly tips back to food! Our animal is doomed to journey back and forth, getting hungrier and thirstier. Each action only equalizes ever-growing urgencies.*

This would be no problem at a dinner table, where food and drink are both within easy reach. But under natural conditions, no animal could survive the waste of energy, when every minor fluctuation caused a major change in strategy. One way to manage this would be to use that "marketplace" infrequently—but that would make our animal less capable of dealing with emergencies. Another way is to use an arrangement called *cross-exclusion*, which appears in many portions of the brain. In such a system, each member of a group of agents is wired to send "inhibitory" signals to all the other agents of that group. This makes them competitors. When any agent of such a group is aroused, its signals tend to inhibit the others. This leads to an avalanche effect—as each competitor grows weaker, its ability to inhibit its challengers also weakens. The result is that even if the initial difference between competitors is small, the most active agent will quickly "lock out" all the others.



Two Examples of Cross-Exclusion

Cross-exclusion arrangements could provide a basis for the principle of "noncompromise" in regions of the brain where competitive mental agents lie close together. Cross-exclusion groups can also be used to construct short-term memory-units. Whenever we force one agent of such a group into activity, even for a moment, it will remain active (and the others will remain suppressed) until the situation is changed by some other strong external influence. Weaker external signals will have scarcely any effect at all because of resistance from within. Why call this a *short-term* memory if it can persist indefinitely? Because when it *does* get changed, no trace will remain of its previous state.

20.1 AMBIGUITY

We often find it hard to "express our thoughts"—to summarize our mental states or put our ideas into words. It is tempting to blame this on the ambiguity of words, but the problem is deeper than that.

Thoughts themselves are ambiguous!

At first, one might complain that that's impossible. *"I'm thinking exactly what I'm thinking; there's no way it could be otherwise. And this has nothing to do with whether I can express it precisely."* But *"what you're thinking now"* is itself inherently ambiguous. If we interpret it to mean the states of *all* your agencies, that would include much that cannot be "expressed" simply because it is not accessible to your language-agency. A more modest interpretation of *"what you're thinking now"* would be a partial indication of the present states of some of your higher-level agencies. But the significance of any agency's state depends on how it is likely to affect the states of other agencies. This implies that in order to "express" your present state of mind, you have to partially anticipate what some of your agencies are about to do. Inevitably, by the time you've managed to express yourself, you're no longer in the state you were before; your thoughts were ambiguous to begin with, and you never *did* succeed in expressing them but merely replaced them with other thoughts.

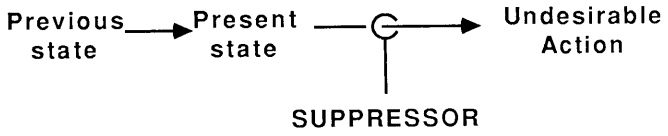
This is not just a matter of words. The problem is that our states of mind are usually subject to change. The properties of physical things tend to persist when their contexts are changed—but the "significance" of a thought, idea, or partial state of mind depends upon which other thoughts are active at the time and upon what eventually emerges from the conflicts and negotiations among one's agencies. It is an illusion to assume a clear and absolute distinction between "expressing" and "thinking," since expressing is itself an active process that involves simplifying and reconstituting a mental state by detaching it from the more diffuse and variable parts of its context.

The listener, too, must deal with ambiguity. You understand *"I wrote a note to my sister,"* despite the fact that the word "note" could mean a short letter or comment, a banknote, a musical sound, an observation, a distinction, or a notoriety. If all our separate words are ambiguous by themselves, why are sentences so clearly understood? Because the context of each separate word is sharpened by the other words, as well as by the context of the listener's recent past. We can tolerate the ambiguity of words because we are already so competent at coping with the ambiguity of thoughts.

27.2 SUPPRESSORS

It would be wonderful never to make mistakes. One way would be to always have such perfect thoughts that none of them is ever wrong. But such perfection can't be reached. Instead we try, as best we can, to recognize our bad ideas before they do much harm. We can thus imagine two poles of self-improvement. On one side we try to stretch the range of the ideas we generate: this leads to more ideas, but also to more mistakes. On the other side, we try to learn not to repeat mistakes we've made before. All communities evolve some prohibitions and taboos to tell their members what they shouldn't do. That, too, must happen in our minds: we accumulate memories to tell ourselves what we shouldn't *think*.

But how could we make an agent to prevent us from doing something that, in the past, has led to bad or ineffectual results? Ideally, that agent would keep us from even *thinking* that bad idea again. But that seems almost paradoxical, like telling someone, "*Don't think about a monkey!*" Yet there is a way to accomplish this. To see how it works, imagine the sequence of mental states that led to a certain mistake:



We could prevent the undesired action from taking place by introducing an agent that recognizes the state which, in the past, preceded the undesired action.

> **Suppressor-agents** *wait until you get a certain "bad idea." Then they prevent your taking the corresponding action, and make you wait until you think of some alternative. If a suppressor could speak, it would say, "Stop thinking that!"*

Suppressors could indeed prevent us from repeating actions that we've learned are bad. But it is inefficient to wait until we actually reach undesirable states, then have to "backtrack." It would be more efficient to anticipate such lines of thought so that we never reach those states at all. In the next section we'll see how to do this by using agents called censors.

> **Censor-agents** *need not wait until a certain bad idea occurs; instead, they intercept the states of mind that usually* **precede** *that thought. If a censor could speak, it would say, "Don't even begin to think that!"*

Though censors were conceived of long ago by Sigmund Freud, they're scarcely mentioned in present-day psychology. I suspect that this is a serious oversight and that censors play fundamental roles in how we learn and how we think. Perhaps the trouble is that our censors work too well. For, naturally, it is easier for psychologists to study only what someone *does*—instead of what someone *doesn't do*.