T HIS TEXT tries to explain how minds work. How can intelligence emerge from non-intelligence? To answer that, I'll show that you can build a mind from many little parts, each mindless by itself.

I call this scheme "Society of Mind," in which each mind is made of many smaller processes. These we'll call agents. Each 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 text. 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 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.

Common sense is not a simple thing. Instead, it is an immense society of many painfully-acquired practical ideas.

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 our infancy, when we formed our first abilities. As each new group of skills matures to work efficiently, we build yet more on top of them, not knowing much of how the old ones work. As time goes on, the layers underneath become increasingly remote till, when we try to speak of them in later life, we find ourselves with little more than synonyms for "I don't know."

One ought to ask, if thinking is so complicated, then how could it all seem so simple? Is it really possible that our minds use such intricate machinery and yet are unaware of it? It scarcely could be otherwise.
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 systems we call "consciousness." Because of this, we cannot trust our offhand judgments about which of the things we do are simple and which of them require complicated machinery. Each portion of the mind can only sense how quietly the other portions do their jobs.

The Investment Principle: Our oldest ideas have unfair advantages over those which come later. The earlier we learn a skill, the more methods we can acquire for using it. Each new idea must then compete, all unprepared, against the mass of skills the old ideas have accumulated.

This is why it's almost always easier to do new things in older ways, instead of starting fresh. Each new idea, however good in principle, seems alien and awkward till we master it. The old ideas keep gaining strength; the new ones rarely can catch up. And then those older, well-developed skills will tend to spread to other worlds of thought in which they do not really work so well — but are just good enough to keep better, new ones from forming at all. In the short run, it is usually easier and better to use bad, old ideas than to start afresh.

Isn't it amazing that a mind can grow, yet keep on working while it's making changes in itself? How do we do it? A mind can't simply shut down work and say it's "closed for renovations." How can we stay alive and well while vital parts are being modified? One way human institutions cope with change is by assigning managers to be responsible for getting important jobs done. And we expect those managers to maintain flexibility by keeping open good alternatives by making several workers interchangeable. Then, while one is being changed or trained, another can be made to do its work. If this is done at every stage, then we can even change the managers. This is one reason we build pyramids of power and authority, both outside and in the mind.

It is an old idea that brains are made of opposing hierarchies. The ascending system must compress large amounts of low-level information into simpler, more meaningful representations, so that those million-featured pictures on our retinas can lead to nemes [thought units] for apples and chairs. The descending system must convert terse instructions from higher levels into multitudes of more specific signals for smaller agents, as when your wish to walk across a room must cause a hundred muscles to pull in a hundred different ways.

People often ask, "Could a machine ever be conscious?" I'm often tempted to ask back, "Could a person ever be conscious?" I mean this as a serious reply, because we seem so ill-equipped to understand ourselves. That in itself is understandable: most of our evolution came long before our brains became intelligent enough to start to know themselves. They still don't seem to have good ways to reach the records of their own activities.

How much genuine self-insight is possible for us? I'm sure our memory machinery provides some useful clues, if only we could learn to interpret them. But it is unlikely that one part of the mind could ever obtain complete descriptions of what happens in the other parts because, it seems, our memory-control systems have too little temporary memory even to represent their own activities in very much detail.

To "know thyself" more perfectly might seem to promise something powerful and good. But there are fallacies concealed behind the happy thought. No doubt, a mind that wants to change itself could benefit from knowing how it works. But then, such knowledge might as easily encourage us to wreck ourselves — if we had ways to poke our clumsy mental fingers into the tricky circuits of the mind's machinery. Could this be why our brains force us to play those games of mental hide and seek?

Just see how prone we are to risk experiments that change ourselves; we're drawn irresistibly to drugs, to meditation, music, even conversation — all powerful addictions that can change our very personalities. Just see how everyone's entranced by any promise to transgress the bounds of normal pleasure and reward.

In ordinary life, our pleasure systems help us learn — and, therefore, to behave ourselves — by forcing checks and balances on us. When "enough is enough," they saturate and satiate.

But when we seize control of them through perverse tricks that break those bounds, then we can reproduce the pleasure of success, yet freed from any need for socially approved accomplishment. And that's the end of everything.

Then, what prevents such meddling? Our minds are bound by many sorts of self-constraint. For one, it's hard to see inside the mind. And even if our inner eyes could see what's there, I think we'd find that many of the agents we'd most want to change would be among the harder ones to change,
the ones which first, in infancy, helped form and shape our longest-lasting self-ideals.

The reason these are hard to change comes from their special evolutionary origin. For the long-term stability of many other mental agencies depends on a certain sluggishness of our images of what we ought to be like. Few of us would survive if, left to random chance, our most adventurous impulses could freely tamper with the basis of our personalities.

In real life, you often have to deal with things you don't completely understand. You drive a car, not knowing how its engine works. You ride as passenger in someone else's car, not knowing how that driver works. Most strange of all, you drive yourself to where you work, not knowing how you work, yourself.

But how do we ever understand anything, really? Almost always, I think, by using one or another kind of analogy. And what is that but to pretend that each new and alien thing we see resembles something we already know. Whenever a new thing's internal workings are too strange, complicated, or unknown to deal with directly, we extract whatever parts of its behavior we can comprehend and represent them by familiar symbols — that is, in terms of familiar things that seem to act in similar ways. This way, we make each novelty appear to be like something we've known before. It is a great idea, that use of words and symbols, icons, images, and names. They let our minds transform the strange into the commonplace.

This then is the point of consciousness: it is a part of the mind that is specialized for knowing how to use other systems which lie hidden in the mind. But it is not a specialist in knowing what those systems actually do, inside themselves. Thus one walks without much sense of how it's really done. It's only when those systems start to fail to work so well that consciousness becomes engaged with small details. That way, a person who has sustained an injured leg may start, for the first time, consciously to make theories about how walking works: "To turn to the left, I'll have to push myself that way" — and then one has to figure out, with what? Actually, we do not often reflect on how our minds solve their problems. I suspect that it is mainly in those moments when we recognize that we're confused that we call up what little knowledge we have about our strategies of thought. Then we find ourselves saying things like this:

"Now I must get organized. Why can't I concentrate on the important questions and not get distracted by those other inessential details?"

Paradoxically, it is very smart to realize that one is confused — that is, in contrast to being confused without knowing it. For then we can apply all our intellect to altering or repairing the defective process.

If our internal mental agents can't communicate, how is it that people can, in spite of having different backgrounds, thoughts, and purposes? The answer is that it is easier for people than for mental agents, because each person knows much more than any smaller portion of that person's mind. Besides, we overestimate how well we actually communicate. We may seem very different from one another in many regards, yet many of our concerns are based on common knowledge and similar experience. This means that we do not really need to tell each other as much as we suppose. Often, when we "explain" something, we merely show some examples of what we mean, and some non-examples; these indicate to the listener how to link up various structures already known. In short, we often just tell "which" kind of thoughts to think, but not "how" to do it.

It would be wonderful to never make mistakes. One way would be to generate such perfect thoughts that none of them are ever, wrong. But such perfection can't be reached. Instead we can try, as best we can, to recognize our bad ideas before they do much harm. So we can imagine two poles of self-improvement. On one side we can try to stretch the range of the ideas we generate: that leads to more ideas, but also to more mistakes. Then, on the other side, we try to learn not to repeat mistakes we've made before. We know that all societies evolve prohibitions and taboos to tell their members what they shouldn't do. That, too, must happen in our minds: we build up banks of memories to tell us what we shouldn't think.

Serious learning tends to change the ways we reach our goals. Humorous learning tends to change the goals themselves.

Our culture regards humor as a pleasant but pointless luxury we do not really need, a thing detached from practicality. Actually, humor has an important, practical function in helping us learn to suppress certain ways of thinking. It censors thoughts.

Humor as a censor would explain why humor is so often concerned with prohibitions and mistakes. Our most productive kinds of thinking are precisely the ones that are most liable to error. Careful, logical thinking can sometimes be made relatively error-free, but then it rarely leads to powerful new ideas. Much more can come from working with analogies and metaphors. The problem is, analogies are often wrong, and metaphors can easily mislead. That's why so many jokes are based on recognizing comparisons that are inept or inappropriate.

Why is it so hard to see that humor plays such
vital roles? Because, I think, it has a funny side effect: when humor turns off other thoughts, it also shuts off thoughts about itself — and thus becomes invisible.

We like to think of memories as though they could restore the things we've known in the past. But memories can't really bring things back; they only reproduce some fragments of our former states of mind, when various sights, sounds, touches, smells, and tastes affected us. But then, what makes some recollections seem so real? The secret is that real-time experience is just as indirect! The closest we can come to apprehending the world, in any case, is through descriptions which our agents make. In fact, if we inquire, instead, about why real things seem real, we'll see that this depends, as well, on memories of things we've known before!

For instance, when you see a telephone, you have a sense, not only of the aspects you can see — its color, texture, size, and shape — but also how it feels to hold it to your ear. You also seem to know at once what telephones are for: that you speak into it here, listen there; that when it rings you answer it; that when you want to call, you dial it. You have a sense of what it weighs, of whether it is soft or hard, of what its other side is like — although you haven't even touched it yet. Those apprehensions come from memories.

Whenever you "get a good idea" or solve a problem or have a memorable experience, you activate a K-line to "represent" it. A K-line is a wirelike agent that attaches itself to whatever mental agents are active when you solve a problem or have a good idea.

When you activate that K-line later, you arouse the agents attached to it, pulling you into a "mental state" much like the one you were in when you solved that problem or got that idea. Because so many of the same agents are active again, you should now find it easier to solve similar kinds of problems!

In other words, we "memorize" what we're thinking about by making a list of the agents involved in that activity. Making a K-line is like making a list of the people who came to a successful party. Here is another image of how K-lines work, suggested by Kenneth Haase, an MIT student who had a large influence on this theory.

"You want to repair a bicycle. Before you start, smear your hands with red paint. Then each tool you need to use will end up with red marks on it. When you're done, just remember that red means "good for fixing bicycles"! Next time you fix a bicycle, you can save time by taking out all the red-marked tools in advance.

"If you use different colors for different kinds of jobs, some tools will end up marked with several different colors. That is, each agent can become attached to many different K-lines. Later, when there's a job to do, just activate the proper K-line for that kind of job, and all the tools used in the past for similar jobs will magically become available."

This suggests a way to make our machine learn to do this job more quickly and easily: we can build a memory that simply keeps a record of which lower-level action-agents were activated during the "Trans-action." So, when we play that sequence back, this "trans-script" would put a second apple in that pail — without invoking any higher-level agencies at all.

We could call this "learning by rote" or doing things so automatically that one can think of other things at the same time. When such a script can work at all, it can work very fast because it has no bureaucracy. But rote-learned skills have serious limitations. They are inflexible because they can work only in narrow ranges of conditions — precisely because they lack bureaucratic
superstructure. They have no higher-level anchor points to use to call for help when anything goes wrong. Because of the lack of hierarchy, there is no place to appeal when an agent gets into trouble.

When someone says "John threw a ball," you probably unconsciously assume some certain set of qualities of color, size and weight. These are your assumptions by default. They might be derived from some ball you owned long ago or, possibly, your newest one. But since such optional details are usually too weak to hold against the sharp insistence of reality, other stimuli will find them easy to detach or otherwise adapt. Defaults don't make strong images, and when they turn out wrong, we aren't too surprised.

But why use defaults at all, instead of simply seeing what's really there? Because, unless we made assumptions, the world would simply make no sense. It would be as useless to perceive how things "actually look" as it would be to watch the random dots on untuned television screens.

What really matters is being able to see "what things look like" — and this is why our brains must have special machinery for representing what we see in terms of artificially distinct "objects" or "things." For the very idea of an object embodies many assumptions we make that "go without saying" — that an object has substance and boundaries, that it existed before we saw it, and that it will remain afterwards — in short, that it will act like most other things. For example, we never see all sides of an object at once, yet we always assume that their unseen sides exist. Perhaps the larger part of what we know is represented by default assumptions, since there is so little we know with perfect certainty.

We use default assumptions in personal relations, too. How does the writer's craft evoke such lifelike characters? It's ridiculous to think that people could be well-portrayed in so few words. Instead, our writers use phrases which activate great networks of assumptions that lie already in their readers' minds. It takes great skill to create those illusions — to activate unknown processes in unknown readers' minds and to shape them to the writer's purposes. Indeed, in doing so, a writer can surpass reality. For, although words are merely catalysts for starting mental processes, so, too, are real things: we can't sense what they really are, but only what we see them to be.

It is not only a matter of language, this ability to simplify or encapsulate other mental processes. We also use this ability to "conceptualize" — that is, to treat ideas as though they were objects — in other areas of thought. Suppose one manages to solve a hard problem after a long and painful search. If we can apply our powers to treat the steps of what we did as though they were parts of an object, then we can "think" about what we did and reassemble the parts that "helped" into a new structure which will do what the old one did, except with much more speed and with much less thought.

An explanation of the difference between older and younger children was first proposed by Seymour Papert in the 1960s, when we first started to explore "Society of Mind" ideas. Most previous theories had tried to explain Piaget's experiments by suggesting that children grow new, different kinds of reasoning as time goes by. That certainly is true, but the importance of Papert's conception is in emphasizing that it is not merely the ingredients of reasoning that matter, but how they're organized: a mind cannot really grow very much by only accumulating more and more new knowledge. It must also develop new and better ways to use what it already knows. That principle deserves a name.

Papert's Principle: Some of the most crucial steps in mental growth are based not simply on acquiring new skills but on acquiring new administrative ways to use what one already knows.

Somehow, each child learns better ways to learn. But it is very hard to guess the nature of those strategies only from watching what the child does. The problem is that those "learning-learning" strategies are twice removed from behavior. If it is hard to guess how our A-brains produce our observable behavior, it must be more than twice as hard to guess how our B-agencies learn to train our A-agencies! Perhaps, in order for a child to become unusually smart, its B-brain must experience a "lucky accident" — the kind that focuses a hidden but persistent interest upon the process of learning itself. It so, then perhaps a major concern for education should be to find out how to get children less concerned with solving particular external problems and more involved in learning better how to learn.

Imagine a simple brain composed of separate "proto-specialists," each concerned with some important need, goal, or instinct like food.

---

There is one way for a mind to watch itself, and yet still contemplate 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.
drink, shelter, reproduction, or defense. On one side, there must be administrators to resolve conflicts between those separate specialists, so that they can work together to control a single body without becoming engaged in paralyzing conflicts. On the other side, each specialist needs ways to use the knowledge that the others gain to fuse the system together. In animals with limited abilities to learn, a loosely-knitted league of almost separate agencies with almost independent goals might be enough to survive in a suitable environment.

But human minds don’t merely learn new ways to reach old goals: they also learn new ways to learn new goals. If we did that without constraint, we’d soon fall prey to accidents — both in the world and in the mind. At the simplest levels, we need protection against accidents like learning not to breathe; at higher levels we must not acquire lethal goals like learning to suppress our other goals entirely — the way that certain saints and mystics do. What sorts of built-in self-constraints could guide a mind toward goals that won’t destroy itself?

No possible inheritance of built-in genes can tell us humans what is good for us! For, unlike all other animals, we build for ourselves the worlds of problems that we face. Then what could teach us what is “good,” if our values change from each generation to the next? The answer is that our goals must be constrained, not by our genes, but by our social heritage — and that’s exactly how we work. Unlike the other animals, we each must learn our goals anew, from one another, through our traditions and cultures. But still the question remains: what mechanism leads us to do that? How could machines that are built by genes help transfer sociocultural goals — without the least idea of what they are? This is done in indirect but specific ways — by exploiting what we call affection, attachment, and love.

Thus our earliest emotions have roots in the machinery through which our original, inborn proto-specialists control what happens in our brains. But soon our cultural surroundings begin to work to override those built-in schemes and start to try to teach us what we ought to feel. First our parents, then our teachers and friends, and finally our own self-ideal twists to impose on us their rules for how to use the mixed-up remnants of those early states; they teach us how and when to let each kind of emotion-sign show. By the time we’re adults, our own expression-systems have become too complicated even for ourselves to understand. What’s more, once we have passed through all the stages of childhood, our grown-up minds have been rebuilt too many times to clearly remember or understand what infants feel. We’re too far removed to be sure that our sympathies, however strong, are authentic.

Yet how could all these steps and stages lead to any sense of unity? Why wouldn’t they, instead, lead us to feel more fragmentary and dispersed? I think the secret is that when each new stage’s work is done, its structure still remains available for further use. Then those remnants of our previous selves supply our personalities with a powerful resource: for whenever one’s present mind becomes confused, it can exploit those stored-up, earlier minds to try old ways to manage things. Although those older selves may not have been so smart, they had much more experience and found many useful ways to cope.

Yet, on the whole, the present personality is almost wholly unaware of this; it has no sense of what it owes to older personalities, because it cannot share their conscious thoughts. And so we just imagine that we have an everpresent Self — a sort of ghostly person-friend inside the head, whom we can always ask for help.

What could be the biological and psychological purposes of the complex, unconscious self-images which grow inside our child-minds?

The answer seems quite clear to me: that’s how Selves start! Consider that our models of ourselves are so complex that even adults can’t explain them. Surely no fragmentary infant mind could know how to build such a complicated thing without some model to base it on. We aren’t born with built-in Selves — but, fortunately, we do arrive with built-in human caretakers. And then, our ancient attachment mechanisms force us to focus on our parents’ ways, to construct crude images of what they want us to be like. Then, stage by stage, those simple models grow until they lead powerful, coherent policies of thought.

This is how the values and goals of a culture are passed from each generation to the next. It is not the same as the “ordinary” kinds of teaching based on signals representing failure or success. Instead, our children learn their deepest values under the influence of attachment-related signals. This is why, when we maintain our standards, we feel virtuous rather than merely successful and
why when we violate those standards we feel shame and guilt rather than mere disappointment. These kinds of emotion are not the same, because they're wired differently.

Why should mental growth proceed by anything like steps at all? Why can't we grow by steady, smooth development?

There are many reasons why it is easier to construct large, complicated things in separated episodes. They are much the same reasons why all large corporations have divisions and departments, and why all complicated living things are divided into organs.

Splitting structures into parts makes them easier to build, maintain and change.

It is always a dangerous experiment to make changes in things that work. What if the new way seems better at first, but later shows some serious flaws — can we get back to where we were? One way might be to keep such complete records that we could “undo” all the changes we've made. But what if those changes caused our quality of thought to become so poor that we could no longer recognize how poor it had become? A safer way would be to keep that older mind intact and build a new one close to it. Then we can use that “previous stage” not only as a “backup” to use if the new one fails, but also to evaluate the new one's performance.

Indeed, one very conservative strategy would be not to permit the new stage actually to take control until it demonstrates the ability to outperform its predecessor. In that case, an outside observer would see “plateaus” followed by sudden spurts of growth. Yet that could be illusory, for the “silent period” might actually conceal the new mind's time of fastest growth. The best part of this scheme is that it permits the person to continue to function while growing — to maintain “business during renovations.” For the “working version” can hold still, while the new one catches up and gets ahead.

It can make sense to think that that there exists, inside your brain, a society of different minds. Like members of a family, they can work together to help each other, each still having its own mental experiences that the others never know about. Several such agencies could have many agents in common, yet still have no more sense of each other's interior activities than people whose apartments share opposite sides of the same walls. Like tenants in a rooming house, the processes that share your brain need not share one another's mental lives.

If each of us contains several such mini-minds, could any special exercise help put them all “in closer touch”? Certainly there are ways to become selectively aware of processes which are not usually conscious at all. But as for becoming aware of everything that happens in one's mind, that surely would leave no room for thought.

As far as I'm concerned, the so-called problem of “body and mind” does not hold any mystery at all:

Minds are simply what brains do.

Whenever we speak about a mind, we're simply speaking of the processes by which our brains proceed from each state to the next state. And this is why minds seem so separate from their physical embodiments: it is simply because nothing can affect that state-succession except the connections that govern how each agent changes its state — and thereby causes other agents to change their states.

In earlier times, we could usually judge machines and process by how they transformed raw materials into finished products. But it makes no sense to speak of brains as though they manufacture thoughts the way factories make cars. The difference is that brains use processes that change themselves — and this means that we cannot separate such processes from the products they produce. In particular, our brains are usually engaged in making memories, which change the ways we'll subsequently think.

Because of this, we cannot trust our common-sense judgments about "thinking machines." Even our technical, scientific theories about such matters are still embryonic, since the whole idea of self-modifying processes is relatively new to science.

The principal activities of brains are making changes in themselves.