Also by Marvin Minsky

THE SOCIETY OF MIND

THE EMOTION MACHINE

Commonsense Thinking,
Artificial Intelligence, and the
Future of the Human Mind

MARVIN MINSKY

Simon & Schuster
New York London Toronto Sydney

Also by Marvin Minsky

THE SOCIETY OF MIND

THE EMOTION MACHINE

Commonsense Thinking,
Artificial Intelligence, and the
Future of the Human Mind

MARVIN MINSKY

Simon & Schuster
New York London Toronto Sydney

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

Verse on page 298 reprinted courtesy of Theodore Melnechuk

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

Designed by Nancy Singer Olaguera

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Minsky, Marvin Lee.
The emotion machine : commonsense thinking,
artificial intelligence, and the future of the human
mind / Marvin Minsky.
p. cm.
Includes bibliographical references (p.) and index.
1. Human information processing. 2. Emotions and cognition. I. Title.

BF444.M56 2006
153—dc22 2006044367

ISBN-13: 978-0-7432-7663-4
ISBN-10: 0-7432-7663-9

To
Gloria, Margaret, Henry, and Juliana

Collaborators
Push Singh
Seymour Papert
John McCarthy
Oliver Selfridge
R. J. Solomonoff

Imprimers
Andrew M. Gleason
George A. Miller
J. C. R. Licklider
Solomon Lefschetz
Warren S. McCulloch
Claude E. Shannon

Supporters
Jeffrey Epstein
Kazuhiko Nishi
Nicholas Negroponte
Harvard Society of Fellows
Office of Naval Research
Toshiba Corporation

CONTENTS

THE EMOTION MACHINE

INTRODUCTION

Nora Joyce, to her husband James:
"Why don't you write books people can read?"

I hope this book will be useful to everyone who seeks ideas about how human minds might work, or who wants suggestions about better ways to think, or who aims toward building smarter machines. It should be useful to readers who want to learn about the field of Artificial Intelligence. It should also be of interest to psychologists, neurologists, computer scientists, and philosophers because it develops many new ideas about the subjects those specialists struggle with.

We all admire great accomplishments in the sciences, arts, and humanities—but we rarely acknowledge how much we achieve in the course of our everyday lives. We recognize the things we see, we understand the words we hear, and we remember things that we've experienced so that, later, we can apply what we've learned to other kinds of problems and opportunities.

We also do a remarkable thing that no other creatures seem able to do: whenever our usual ways to think fail, *we can start to think about our thoughts themselves*—and if this "reflective thinking" shows where we went wrong, that can help us to invent new and more powerful ways to think. However, we still know very little about how our brains manage to do such things. How does imagination work? What are the causes of consciousness? What are emotions, feelings, and thoughts? How do we manage to think at all?

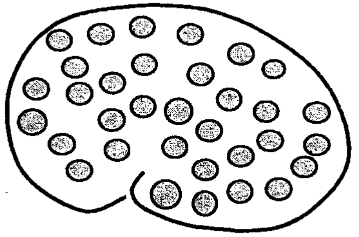
Contrast this with the progress we've seen toward answering questions about physical things. What are solids, liquids, and gases? What are colors, sounds, and temperatures? What are forces, stresses, and strains? What is the nature of energy? Today, almost all such mysteries have been explained in terms of very small numbers of simple laws—for example, the equations discovered by such physicists as Newton, Maxwell, Einstein, and Schrödinger.

So naturally, psychologists tried to imitate physicists—by searching for compact sets of laws to explain what happens inside our brains. However, no such simple set of laws exists, because every brain has hundreds of parts, each of which evolved to do certain particular kinds of jobs; some of them recognize situations, others tell muscles to execute actions, others formulate goals and plans, and yet others accumulate and use enormous bodies of knowledge. And though we don't yet know enough about how each of those brain-centers works, we do know their construction is based on information that is contained in tens of thousands of inherited genes, so that each brain-part works in a way that depends on a somewhat different set of laws.

Once we recognize that our brains contain such complicated machinery, this suggests that we need to do the opposite of what those physicists did: instead of searching for simple explanations, we need to find more complicated ways to explain our most familiar mental events. The meanings of words like "feelings," "emotions," or "consciousness" seem so natural, clear, and direct to us that we cannot see how to start thinking about them. However, this book will argue that none of those popular psychology words refers to any single, definite process; instead each of those words attempts to describe the effects of large networks of processes inside our brains. For example, Chapter 4 will demonstrate that "consciousness" refers to more than twenty different such processes!

It might appear to make everything worse, to change some things that looked simple at first into problems that now seem more difficult. However, on a larger scale, this increase in complexity will actually make our job easier. For, once we split each old mystery into parts, we will have replaced each old, big problem with several new and smaller ones—each of which may still be hard but no longer will seem unsolvable. Furthermore, Chapter 9 will argue that regarding ourselves as complex machines need not diminish our feelings of self-respect, and should enhance our sense of responsibility.

To start dividing those old big questions into smaller ones, this book will begin by portraying a typical brain as containing a great many parts that we'll call "resources."



We'll use this image whenever we want to explain some mental activity (such as Anger, Love, or Embarrassment) by trying to show how that state of mind might result from the activities of a certain collection of mental resources. For example, the state called "Anger" appears to arouse resources that make us react with unusual speed and strength—while suppressing resources that we otherwise use to plan and act more prudently; thus, Anger replaces your cautiousness with aggressiveness and trades your sympathy for hostility. Similarly, the condition called "Fear" would engage resources in ways that cause you to retreat.

> Citizen: I sometimes find myself in a state where everything seems cheerful and bright. Other times (although nothing has changed) all my surroundings seem dreary and dark, and my friends describe me as "down" or "depressed." Why do I have such states of mind—or moods, or feelings, or dispositions—and what causes all of their strange effects?

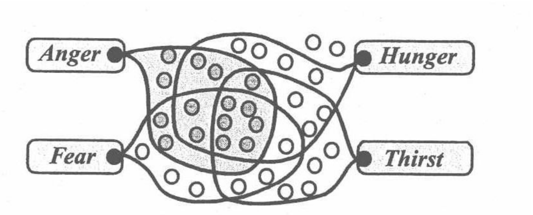
Some popular answers to this are, *"Those changes are caused by chemicals in the brain,"* or *"They result from an excess of stress,"* or *"They come from thinking depressing thoughts."* However, such statements say almost nothing about how those processes actually work—whereas the idea of selecting a set of resources can suggest more specific ways in which our thinking can change. For example, Chapter 1 will begin by thinking about this very familiar phenomenon:

> When a person you know has fallen in love, it's almost as though someone new has emerged—a person who thinks in other ways, with altered goals and priorities. It's almost as though a switch had been thrown and a different program has started to run.

What could happen inside a brain to make such changes in how it thinks? Here is the approach this book will take:

> *Each of our major "emotional states" results from turning certain resources on while turning certain others off—and thus changing some ways that our brains behave.*

But what activates such sets of resources? Our later chapters will argue that our brains must also be equipped with resources that we shall call *"Critics"*—each of which is specialized to recognize some certain condition—and then to activate a specific collection of other resources. Some of our Critics are built in from birth, to provide us with certain "instinctive" reactions—such as Anger, Hunger, Fear, and Thirst—which evolved to help our ancestors survive. Thus, Anger and Fear evolved for defense and protection, while Hunger and Thirst evolved for nutrition.



However, as we learn and grow, we also develop ways to activate other, new sets of resources to use—and this leads to types of mental states that we regard as more "intellectual" than "emotional." For example, whenever a problem seems hard to you, then your mind will start to switch among different Ways to Think—by selecting different sets of resources that can help you to divide the problem into smaller parts, or find suggestive analogies, or retrieve solutions from memories—or even ask some other person for help.

The rest of this book will argue that this could be what provides our species with our uniquely human resourcefulness.

> *Each of our major Ways to Think results from turning certain resources on while turning certain others off—and thus changing some ways that our brains behave.*

For example, our first few chapters will try to show how this could explain such states of mind as Love, Attachment, Grief, and Depression in terms of how they exploit our resources. Then the later chapters will do the same for more "intellectual" sorts of thought.

> Citizen: It seems strange that you've given the same description both for emotions and for regular thinking. But thinking is basically rational—dry, detached, and logical—whereas emotions enliven our ways to think by adding irrational feelings and biases.

There is a traditional view in which emotions *add* extra features to plain, simple thoughts, much as artists use colors to augment the effects of black-and-white drawings. However, this book will argue, instead, that many of our emotional states result when certain particular Ways to Think start to *suppress* our use of certain resources! For example, Chapter 1 will portray "infatuation" as a condition in which we suppress some resources that we might otherwise use to recognize faults in somebody else. Besides, I think it's a myth that there's any such thing as purely logical, rational thinking—because our minds are always affected by our assumptions, values, and purposes.

> Citizen: I still think your view of emotions ignores too much. For example, emotional states like fear and disgust involve the body as well as the brain, as when we feel discomfort in the chest or gut, or palpitations of the heart, or when we feel faint or tremble or sweat.

I agree that this view may seem too extreme—but sometimes, to explore new ideas, we need to set our old ones aside, at least temporarily. For example, in the most popular view, emotions are deeply involved with our bodies' conditions. However, Chapter 7 will take the opposite view, by regarding our body parts as resources that our brains can use to change (or maintain) their mental states! For example, you sometimes can make yourself persist at a plan by maintaining a certain facial expression.

So, although this book is called *The Emotion Machine,* it will argue that emotional states are not especially different from the processes that we call "thinking"; instead, emotions are certain ways to think that we use to increase our resourcefulness—that is, when our passions don't grow till they handicap us—and this variety of Ways to Think must be such a substantial part of what we call "intelligence" that perhaps we should call it "resourcefulness." And this applies not only to emotional states but to all of our mental activities:

> *If you "understand"' something in only one way, then you scarcely understand it at all—because when you get stuck, you'll have nowhere to go. But if you represent something in several ways, then when you get frustrated enough, you can switch among different points of view, until you find one that works for you!*

Accordingly, when we design machines to mimic our minds—that is, to create Artificial Intelligences—we'll need to make sure that those machines, too, are equipped with sufficient diversity:

> If a program works in only one way, then it gets stuck when that method fails. But a program that has several ways to proceed could then switch to some other approach, or search for a suitable substitute.

This idea is a central theme of this book—and it is firmly opposed to the popular view that each person has a central core—some sort of invisible spirit or self—from which all their mental abilities originate. For that seems a demeaning idea—that all our virtues are secondhand—or that we deserve no credit for our accomplishments, because they come to us as gifts from some other source. Instead, I see our dignity as stemming from what we each have made of ourselves: a colossal collection of different ways to deal with different situations and predicaments. It is that diversity that distinguishes us from most of the other animals—and from all the machines that we've built in the past—and every chapter of this book will discuss some of the sources of our uniquely human resourcefulness.

Chapter 1. We are born with many mental resources.
Chapter 2. We learn more from interacting with others.
Chapter 3. Emotions are different Ways to Think.
Chapter 4. We learn to think about our recent thoughts.
Chapter 5. . We learn to think on multiple levels.
Chapter 6. We accumulate huge stores of commonsense knowledge.
Chapter 7. We switch among different Ways to Think.
Chapter 8. We find multiple ways to represent things.
Chapter 9. We build multiple models of ourselves.

For centuries, psychologists searched for ways to explain our everyday mental processes—yet many thinkers still today regard the nature of mind as a mystery. Indeed, it still is widely believed that minds are made of ingredients that can only exist in living things, that no machine could feel or think, worry about what might happen to it, or even be conscious that it exists—or could ever develop the kinds of ideas that could lead to great paintings or symphonies.

This book will pursue all those goals at once: to suggest how human brains might work and to design machines that can feel and think. Then we can try to apply those ideas both to understand ourselves and to develop Artificial Intelligence.

How This Book Handles Quotations and References

Each statement in quotation marks is by an actual person; if it also has a date, the source will be in the bibliography.

> Marcel Proust 1927: "Each reader reads only what is already inside himself. A book is only a sort of optical instrument which the writer offers to let the reader discover in himself what he would not have found without the aid of the book."

A statement without quotation marks is a fictional comment a reader might make.

> Citizen: If our everyday thinking is so complex, then why does it seem so straightforward to us?

Most references are conventional bibliographical citations, such as

> Schank, 1975: Roger C. Schank, *Conceptual Information Processing.* New York: American Elsevier, 1975.

Some references are to pages on the World Wide Web.

> Lenat 1998: Douglas B. Lenat. *The Dimensions of Context Space.* Available at http://www.cyc.com/doc/context-space.pdf.

Some other references are to "newsgroups" on the Web, such as

> McDermott 1992: Drew McDermott, In comp.ai.philosophy. February 7, 1992.

To access such newsgroup documents (along with the context in which they were written) one can make a Google search for comp.ai.philosophy McDermott 1992. So I will try to maintain copies of these on my Web site at www.emotionmachine.net. Readers are also invited to use that site for sending questions and comments to me.

Note: This book uses the term *"resource"* where my earlier book, *The Society of Mind,* used *"agent."* I made this change because too many readers assumed that an "agent" is a personlike thing (like a travel agent) that could operate independently, or cooperate with others in much the same ways that people do. On the contrary, most resources are specialized to certain kinds of jobs for certain other resources, and cannot directly communicate with most of the person's other resources. For more details about how these two books relate, see the article by Push Singh 2003, who helped to develop many of the ideas in this book.

1
FALLING IN LOVE

1-1 Infatuation

> "In faith, I do not love thee with mine eyes,
> For they in thee a thousand errors note;
> But 'tis my heart that loves what they despise."
> —*Shakespeare*

Many people find it absurd to think of a person as like a machine—so we often hear such statements as this:

> Citizen: Of course machines can do useful things. We can make them add up huge columns of numbers or assemble cars in factories. But nothing made of mechanical stuff could ever have genuine feelings like love.

No one finds it surprising these days when we make machines that do logical things, because logic is based on clear, simple rules of the sorts that computers can easily use. But *Love,* by its nature, some people would say, cannot be explained in mechanical ways—nor could we ever make machines that possess any such human capacities as feelings, emotions, and consciousness.

What is Love, and how does it work? Is this something that we want to understand, or is it one of those subjects that we don't really want to know more about? Hear our friend Charles attempt to describe his latest infatuation.

> "I've just fallen in love with a wonderful person. I scarcely can think about anything else. My sweetheart is unbelievably perfect—of indescribable beauty, flawless character, and incredible intelligence. There is nothing I would not do for her."

On the surface such statements seem positive; they're all composed of superlatives. But note that there's something strange about this: most of those phrases of positive praise use syllables like "*un,*" "*less,*" and "*in*"—which show that they really are negative statements describing the person who's saying them!

> Wonderful. Indescribable.
> (I can't figure out what attracts me to her.)
> I scarcely can think of anything else.
> (Most of my mind has stopped working.)
> Unbelievably perfect. Incredible.
> (No sensible person believes such things.)
> She has a flawless character.
> (I've abandoned my critical faculties.)
> There is nothing I would not do for her.
> (I've forsaken most of my usual goals.)

Our friend sees all this as positive. It makes him feel happy and more productive, and relieves his dejection and loneliness. But what if most of those pleasant effects result from his success at suppressing his thoughts about what his sweetheart actually says:

> "Oh, Charles—a woman needs certain things. She needs to be loved, wanted, cherished, sought after, wooed, flattered, cosseted, pampered. She needs sympathy, affection, devotion, understanding, tenderness, infatuation, adulation, idolatry—that isn't much to ask, is it, Charles?"[1]

Thus, Love can make us disregard most defects and deficiencies, and make us deal with blemishes as though they were embellishments—even when, as Shakespeare said, we still may be partly aware of them:

> "When my love swears that she is made of truth,
> I do believe her, though I know she lies."

We are equally apt to deceive ourselves, not only in our personal lives but also when dealing with abstract ideas. There, too, we often close our eyes to conflicts and clashes between our beliefs. Listen to Richard Feynman's words:

> "That was the beginning and the idea seemed so obvious to me that I fell deeply in love with it. And, like falling in love with a woman, it is only possible if you don't know too much about her, so you cannot see her faults. The faults will become apparent later, but after the love is strong enough to hold you to her. So, I was held to this theory, in spite of all the difficulties, by my youthful enthusiasm."
>
> —*1966 Nobel Prize lecture*

What does a lover actually love? That should be the person to whom you're attached—but if your pleasure mainly results from suppressing your other questions and doubts, then you're only in love with Love itself.

> Citizen: So far, you have spoken only about what we call infatuation—sexual lust and extravagant passion. That leaves out most of the usual meanings of "love"—such as tenderness, trust, and companionship.

Indeed, once those short-lived attractions fade, they sometimes go on to be replaced by more enduring relationships, in which we exchange our own interests for those of the persons to whom we're attached:

> Love, n. That disposition or state of feeling with regard to a person which (arising from recognition of attractive qualities, from instincts of natural relationship, or from sympathy) manifests itself in solicitude for the welfare of the object, and usually also in delight in his or her presence and desire for his or her approval; warm affection, attachment.
>
> —Oxford English Dictionary

Yet even this larger conception of Love is still too narrow to cover enough, because *Love* is a kind of suitcase-like word, which includes other kinds of attachments like these:

The love of a parent for a child.
A child's affection for parents and friends.
The bonds that make lifelong companionships.
The connections of members to groups or their leaders.

We also apply that same word Love to our involvements with objects, feelings, ideas, and beliefs—and not only to ones that are sudden and brief, but also to bonds that increase through the years.

A convert's adherence to doctrine or scripture.
A patriot's allegiance to country or nation.
A scientist's passion for finding new truths.
A mathematician's devotion to proofs.

Why do we pack such dissimilar things into those single suitcase-words? As we'll see in Section 1-3, each of our common "emotional" terms describes a variety of different processes. Thus we use the word Anger to abbreviate a diverse collection of mental states, some of which change our ways to perceive, so that innocent gestures get turned into threats—and thus make us more inclined to attack. Fear also affects the ways we react but makes us retreat from dangerous things (as well as from some that might please us too much).

Returning to the meanings of Love, one thing seems common to all those conditions: *each leads us to think in different ways:*

> *When a person you know has fallen in love, it's almost as though someone new has emerged—a person who thinks in other ways, with altered goals and priorities. It's almost as though a switch had been thrown and a different program has started to run.*

This book is mainly filled with ideas about what could happen inside our brains to cause such great changes in how we think.

1-2 The Sea of Mental Mysteries

From time to time we think about how we try to manage our minds:

Why do I waste so much of my time?
What determines to whom I'm attracted?

Why do I have such strange fantasies?
Why do I find mathematics so hard?
Why am I afraid of heights and crowds?
What makes me addicted to exercise?

But we can't hope to understand such things without adequate answers to questions like these:

What sorts of things are emotions and thoughts?
How do our minds build new ideas?
What are the bases for our beliefs?
How do we learn from experience?
How do we manage to reason and think?

In short, we all need better ideas about the ways in which we think. But whenever we start to think about that, we encounter yet more mysteries.

What is the nature of consciousness?
What are feelings and how do they work?
How do our brains imagine things?
How do our bodies relate to our minds?
What forms our values, goals, and ideals?

Now, everyone knows how Anger feels—or Pleasure, Sorrow, Joy, and Grief—yet we still know almost nothing about how those processes actually work. As Alexander Pope asks in his *Essay on Man,* are these things that we can hope to understand?

"Could he, whose rules the rapid comet bind,
Describe or fix one movement of his mind?
Who saw its fires here rise, and there descend,
Explain his own beginning, or his end?"

How did we manage to find out so much about atoms and oceans and planets and stars—yet so little about the mechanics of minds? Thus, Newton discovered just three simple laws that described the motions of all sorts of objects; Maxwell uncovered just four more laws that explained all electromagnetic events; then Einstein reduced all those and more into yet smaller formulas. All this came from the success of those physicists' quest: *to find simple explanations for things that seemed, at first, to be highly complex.*

Then, why did the sciences of the mind make less progress in those same three centuries? I suspect that this was largely because most psychologists mimicked those physicists, by looking for equally compact solutions to questions about mental processes. However, that strategy never found small sets of laws that accounted for, in substantial detail, any large realms of human thought. So this book will embark on the opposite quest: *to find* more complex *ways to depict mental events that seem simple at first!*

This policy may seem absurd to scientists who have been trained to believe such statements as, *"One should never adopt hypotheses that make more assumptions than one needs."* But it is worse to do the opposite—as when we use "psychology words" that mainly hide what they try to describe. Thus, every phrase in the sentence below conceals its subject's complexities:

> *You look at an object and see what it is.*

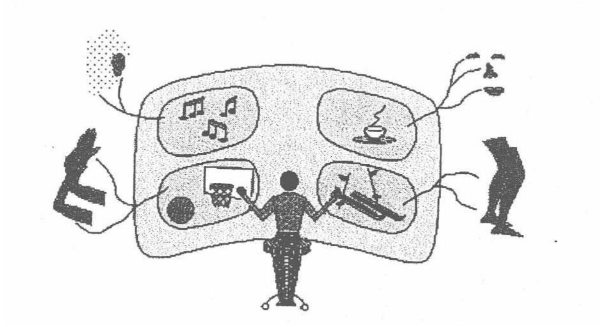
For, *"look at"* suppresses your questions about the systems that choose how you move your eyes. Then, *"object"* diverts you from asking how your visual systems partition a scene into various patches of color and texture—and then assign them to different "things." Similarly, *"see what it is"* serves to keep you from asking how recognitions relate to other things that you've seen in the past.

It is the same for most of the commonsense words we use when we try to describe the events in minds—as when one makes a statement like, *"I think I understood what you said."* Perhaps the most extreme examples of this are when we use words like *you* and *me,* because we all grow up with this fairy tale:

> *We each are constantly being controlled by powerful creatures inside our minds who do our feeling and thinking for us, and make our important decisions for us. We call these our "Selves" or "Identities"—and we believe that they always remain the same, no matter how we may otherwise change.*

This *"Single-Self"* concept serves us well in our everyday social affairs. But it hinders our efforts to think about what minds are and how they work—because, when we ask about what Selves actually do, we get the same answer to every such question:

> *Your Self sees the world by using your senses. Then it stores what it learns in your memory. It originates all your desires and goals—and then solves all your problems for you, by exploiting your "intelligence."*



A Self Controlling Its Person's Mind

What attracts us to this queer idea, that we don't make any decisions ourselves but delegate them to some other entity? Here are a few kinds of reasons why a mind might entertain such a fiction:

> Child psychologist: As a child, you learned to distinguish among some persons in your environment. Later, you somehow came to conclude that you are such a person, too—but at the same time, you may have assumed that there is a person inside of you.

> Psychotherapist: The Single-Self legend helps makes life seem pleasant, by hiding from us how much we're controlled by all sorts of conflicting, unconscious goals.

> Practical person: That image makes us efficient, whereas better ideas might slow us down. It would take too long for our hard-working minds to understand everything all the time.

However, although the Single-Self concept has practical uses, it does not help us to understand ourselves—because it does not provide us with *smaller parts* we could use to build theories of what we are. When you think of yourself as a single thing, this gives you no clues about issues like these:

What determines the subjects I think about?
How do I choose what next to do?
How can I solve this difficult problem?

Instead, the Single-Self concept offers only useless answers like these:

My Self selects what to think about.
My Self decides what I should do next.
I should try to make my Self get to work.

Whenever we wonder about our minds, the simpler are the questions we ask, the harder it seems to find answers to them. When asked about a complex physical task like, *"How could a person build a house,"* you might answer almost instantly, *"Make a foundation and then build walls and a roof."* However, we find it much harder to think of what to say about seemingly simpler questions like these:

How do you recognize things that you see?
How do you comprehend what a word means?
What makes you like pleasure more than pain?

Of course, those questions are not really simple at all. To "see" an object or "speak" a word involves hundreds of different parts of your brain, each of which does some quite difficult jobs. Then why don't we sense that complexity? That's because most such jobs are done inside parts of the brain whose internal processes are hidden from the rest of the brain.

At the end of this book, we'll come back to examine the concepts of Self and Identity, and conclude that the structures that we call our Selves are elaborate structures that each of us builds to use for many purposes.

Whenever you think about your "Self," you are switching among a huge network of models, each of which tries to represent some particular aspects of your mind—to answer some questions about yourself.

1-3 Moods and Emotions

William James 1890: "If one should seek to name each particular one of them of which the human heart is the seat, each race of men having found names for some shade of feeling which other races have left undiscriminated . . . all sorts of groupings would be possible, according as we chose this character or that as a basis. The only question would be, does this grouping or that suit our purpose best?"

Sometimes a person gets into a state where everything seems to be cheerful and bright—although nothing outside has actually changed. Other times everything pleases you less: the entire world seems dreary and dark, and your friends complain that you seem depressed. Why do we have such states of mind—or moods, or feelings, or dispositions—and what causes all their strange effects? Here are some of the phrases we find when dictionaries define *emotion.*

The subjective experience of a strong feeling.
A state of mental agitation or disturbance.
A mental reaction involving the state of one's body.
A subjective rather than conscious affection.
The parts of consciousness that involve feeling.
A nonrational aspect of reasoning.

If you didn't yet know what emotions are, you certainly wouldn't learn much from this. What is *subjective* supposed to mean, and what could a *conscious affection* be? In what ways do those *parts of consciousness* become *involved* with what we call *"feelings"*? Must every emotion involve a *disturbance*? Why do so many such questions arise when we try to define what *emotion* means?

The reason for this is simply that *emotion* is one of those suitcaselike words that we use to conceal the complexity of very large ranges of different things whose relationships we don't yet comprehend. Here are a few of the hundreds of terms that we use to refer to our mental conditions:

Admiration, Affection, Aggression, Agitation, Agony, Alarm, Ambition, Amusement, Anger, Anguish, Anxiety, Apathy, Assurance, Attraction, Aversion, Awe, Bliss, Boldness, Boredom, Confidence, Confusion, Craving, Credulity, Curiosity, Dejection, Delight, Depression, Derision, Desire, Detest, Disgust, Dismay, Distrust, Doubt, etc.

Whenever you change your mental state, you might try to use those emotion-words to try to describe your new condition—but usually each such word or phrase refers to too wide a range of states. Many researchers have spent their lives at classifying our states of mind, by arranging terms like *feelings, dispositions, tempers,* and *moods* into orderly charts or diagrams—but should we call *Anguish* a feeling or a mood? Is *Sorrow* a type of disposition? No one can settle the use of such terms because different traditions make different distinctions, and different people have different ideas about how to describe their various states of mind. How many readers can claim to know precisely how each of the following feelings feels?[2]

Grieving for a lost child
Fearing that nations will never live in peace
Rejoicing in an election victory
Excited anticipation of a loved one's arrival
Terror as your car loses control at high speed
Joy at watching a child at play
Panic at being in an enclosed space

In everyday life, we expect our friends to know what we mean by *Pleasure* or *Fear*—but I suspect that attempting to make our old words more precise has hindered more than helped us to make theories about how human minds work. So this book will take a different approach, by thinking of each mental condition as based on the use of many small processes.

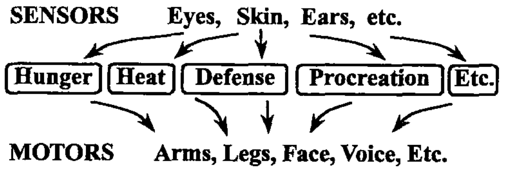
1-4 Infant Emotions

> Charles Darwin 1872: "Infants, when suffering even slight pain, moderate hunger, or discomfort, utter violent and prolonged screams. Whilst thus screaming their eyes are firmly closed, so that the skin round them is wrinkled, and the forehead contracted into a frown. The mouth is widely opened with the lips retracted in a peculiar manner, which causes it to assume a squarish form; the gums or teeth being more or less exposed."

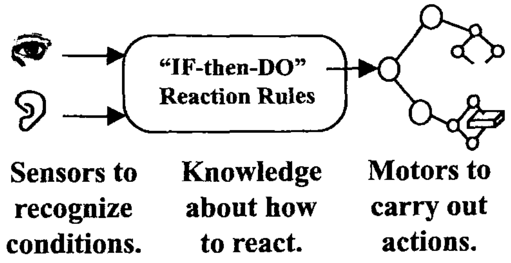
One moment your baby seems perfectly well, but then come some restless motions of limbs. Next you see a few catches of breath, and then suddenly the air fills with screams. Is baby hungry, sleepy, or wet? Whatever the trouble may turn out to be, those cries compel you to find some way to help—and once you find the remedy, things quickly return to normal. In the meantime though, you, too, feel distressed. When a friend of yours cries, you can ask her what's wrong—but when your baby abruptly changes his state, there may seem to be "no one home" to communicate with.

Of course, I do not mean to suggest that infants don't have "personalities." Soon after birth you can usually sense that a particular baby reacts more quickly than others, or seems more patient or irritable, or even more inquisitive. Some of those traits may change with time, but others persist throughout life. Nevertheless, we still need to ask, What could make an infant so suddenly switch, between one moment and the next, from contentment or calmness to anger or rage?

To answer that kind of question, you would need a theory about the machinery that underlies that infant's behavior. So let's imagine that someone has asked you to build an artificial animal. You could start by making a list of goals that your animal-robot needs to achieve. It may need to find parts with which to repair itself. It may need defenses against attacks. Perhaps it should regulate its temperature. It may even need ways to attract helpful friends. Then once you have assembled that list, you could tell your engineers to meet each of those needs by building a separate "instinct-machine"—and then to package them all into a single "body-box."

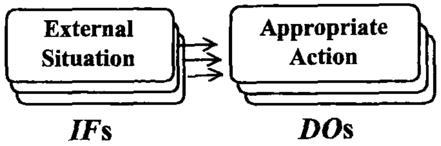


What goes inside each instinct-machine? Each of them needs three kinds of resources: some ways to recognize situations, some knowledge about how to react to these, and some muscles or motors to execute actions.


"IF-then-DO"
Reaction Rules
Sensors to recognize conditions.
Knowledge about how to react.
Motors to carry out actions.

What goes inside that knowledge box? Let's begin with the simplest case: suppose that we already know, in advance, all the situations our robot will face. Then all we need is a catalog of simple, two-part *"If→Do"* rules—where each *If* describes one of those situations, and each *Do* describes an action to take. Let's call this a *"Rule-Based Reaction-Machine."*

Rule Based Reaction-Machine

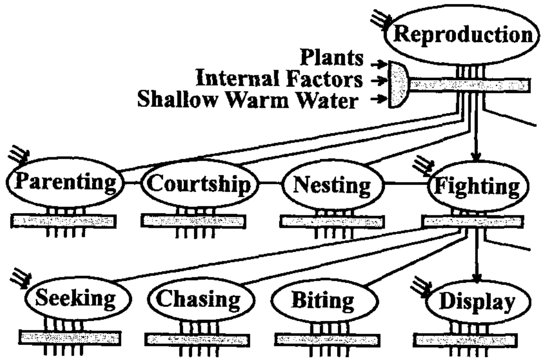


If you are too hot, *Move* into the shade.
If you are hungry, *Find* something to eat.
If you're facing a threat, *Select* some defense.

Every infant animal is born with many *If→Do* rules like these. For example, each human infant is born with ways to maintain its body temperature: when too hot, it can pant, sweat, stretch out, or vasodilate; when too cold, it can shiver, retract its limbs, or vasoconstrict—or metabolize to produce more heat. Then later in life, we learn to use actions that change the external world.

If you are too cold, *Turn* on a heater.
If your room is too hot, *Open* a window.
If there's too much sunlight, *Pull* down the shade.

It would be naive to try to describe a mind as nothing more than bundles of *If→Do* rules. However, the great animal psychologist Nikolaas Tinbergen showed in his book *The Study of Instinct*[3] that when such rules are combined in certain ways, they can account for a remarkable range of different things that animals do. This sketch shows only a part of the structure that Tinbergen proposed to explain how a certain fish behaves.



Of course, it would need much more than this to support the higher levels of human thought. The rest of this book will describe some ideas about the structures inside our human minds.

1-5 Seeing a Mind As a Cloud of Resources

We all know ways to describe our minds, as they appear to us when seen from outside:

> Albert Einstein 1950: "We are all ruled in what we do by impulses; and these impulses are so organized that our actions in general serve for our self preservation and that of the race. Hunger, love, pain, fear are some of those inner forces which rule the individual's instinct for self preservation. At the same time, as social beings, we are moved in the relations with our fellow beings by such feelings as sympathy, pride, hate, need for power, pity, and so on."

This book will try to show how such states of mind could come from machines inside our brains. To be sure, many thinkers still insist that machines can never feel or think.

> Citizen: A machine can do only what it is programmed to do, and does it without any thinking or feeling. No machine can get tired or bored or have any kind of emotion at all. It cannot care when something goes wrong, and, even when it gets things right, it feels no sense of pleasure, pride, or delight in those accomplishments.

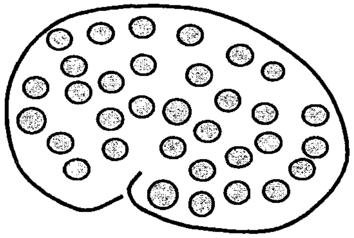
> Vitalist: That's because machines have no spirits or souls, and no wishes, ambitions, desires, or goals. That's why a machine will just stop when it's stuck—whereas a person will struggle to get something done. Surely this must be because people are made of different stuff; we are alive and machines are not.

In earlier times, those views seemed plausible, because living things seemed so different from machines—and no one could even begin to conceive of how physical things could feel or think. But once we developed more scientific instruments (and better ideas about science itself), then "life" became less mysterious, because now we could see that each living cell consists of hundreds of kinds of machinery.

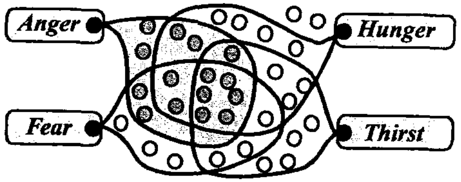
> Holist: Yes, but many people still maintain that there will always remain a mystery about how a living thing could ever result from nothing more than mechanical stuff. Surely we're more than the sum of our parts.

That once was a popular belief, but today it is widely recognized that behavior of a complex machine depends only on how its parts interact, but not on the "stuff" of which they are made (except for matters of speed and strength). In other words, all that matters is the manner in which each part reacts to the other parts to which it is connected. For example, we can build computers that behave in identical ways, no matter if they consist of electronic chips or of wood and paper clips—provided that their parts perform the same processes, so far as the other parts can see.

This suggests replacing old questions like, "What sorts of things are emotions and thoughts?" by more constructive ones like, "What *processes* does each emotion involve?" and "How could machines perform such *processes*?" To do this, we'll start with the simple idea that every brain contains many parts, each of which does certain specialized jobs. Some can recognize various patterns, others can supervise various actions, yet others can formulate goals or plans, and some can contain large bodies of knowledge. This suggests that we could envision a mind (or a brain) as composed of a great many different "resources."



At first this image may seem hopelessly vague—yet it can help us start to understand how a mind could make a large change in its state. For example, the state we call "angry" could be what happens when you activate some resources that help you react with more speed and strength—while also suppressing some other resources that usually make you act prudently. This will replace your usual cautiousness with aggressiveness, change empathy into hostility, and cause you to plan less carefully. All of this could result from turning on the resource labeled Anger in this diagram:



Similarly, we could explain such mental conditions as Hunger and Fear—and we could even account for what happened to Charles in his state of acute infatuation: perhaps such a process turned off the resources he normally used to recognize another person's faults—and also supplanted some of his usual goals by ones that he thought Celia wants him to hold. So now, let's make a generalization:

> *Each of our major "emotional states" results from turning certain resources on while turning certain others off—thus changing the way one's brain behaves.*